TESIS – TE142599

SEGMENTASI PENGGUNA WEB MENGGUNAKAN METODE *GENETIC K-MEANS ALGORITHM*

NUR ULFATUR ROIHA
2214206701

DOSEN PEMBIMBING
Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
Dr. Adhi Dharma Wibawa, ST., MT.

PROGRAM MAGISTER
BIDANG KEAHLIAN TELEMATIKA-CIO
JURUSAN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2017

TESIS – TE142599

SEGMENTASI PENGGUNA WEB MENGGUNAKAN METODE *GENETIC K-MEANS ALGORITHM*

NUR ULFATUR ROIHA
2214206701

DOSEN PEMBIMBING
Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
Dr. Adhi Dharma Wibawa, ST., MT.

PROGRAM MAGISTER
BIDANG KEAHLIAN TELEMATIKA-CIO
JURUSAN TEKNIK ELEKTRO
FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2017

TESIS – TE142599

WEB USERS SEGMENTATION USING GENETIC K-MEANS ALGORITHM METHOD

NUR ULFATUR ROIHA
2214206701

SUPERVISIORS
Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
Dr. Adhi Dharma Wibawa, ST., MT.

MAGISTER PROGRAM
PROGRAM TELEMATIKA-CIO
DEPARTMEN OF ELECTRICAL ENGINEERING
FACULTY OF INDUSTRIAL TECHNOLOGY
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2017

PERNYATAAN KEASLIAN TESIS

Dengan ini saya menyatakan bahwa isi sebagian atau keseluruhan tesis saya dengan judul **"SEGMENTASI PENGGUNA WEB MENGGUNAKAN METODE *GENETIC K-MEANS ALGORITHM*"** adalah benar karya intelektual mandiri, diselesaikan tanpa menggunakan bahan-bahan yang tidak diijinkan dan bukan karya dari pihak lain yang saya akui sebagai karya sendiri.

Semua referensi yang dikutip maupun yang dirujuk telah ditulis secara lengkap pada daftar pustaka. Apabila ternyata pernyataan ini tidak benar, saya bersedia menerima sanksi sesuai peraturan yang berlaku.

Surabaya, Januari 2017

Nur Ulfatur Roiha
NRP. 2214206701

Halaman ini sengaja dikosongkan

LEMBAR PENGESAHAN

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Teknik (M.T)
di
Institut Teknologi Sepuluh Nopember

oleh:

Nur Ulfatur Roiha
NRP. 2214206701

Tanggal Ujian : 06 Januari 2017
Periode Wisuda : Maret 2017

Disetujui oleh:

1. **Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.** **(Pembimbing I)**
   NIP: 195409251978031001

2. **Dr. Adhi Dharma Wibawa, ST., MT.** **(Pembimbing II)**
   NIP: 197605052008121003

3. **Prof. Dr. Ir. Mauridhi Hery Purnomo, M.Eng.** **(Penguji)**
   NIP: 195809161986011001

4. **Dr. Ir. Endroyono, DEA.** **(Penguji)**
   NIP: 196504041991021001

5. **Dr. Surya Sumpeno, ST., M.Sc.** **(Penguji)**
   NIP: 196906131997021003

an. Direktur Program Pascasarjana
Asisten Direktur

Prof. Dr. Ir. Tri Widjaja, M.Eng.
NIP. 196110211986031001

Halaman ini sengaja dikosongkan

SEGMENTASI PENGGUNA WEB MENGGUNAKAN METODE GENETIC K-MEANS ALGORITHM

Nama mahasiswa : Nur Ulfatur Roiha
NRP : 2214206701
Pembimbing : 1. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
2. Dr. Adhi Dharma Wibawa, ST., MT.

ABSTRAK

Kebutuhan dan ketergantungan terhadap internet semakin hari semakin meningkat yang menyebabkan trafik internetpun meningkat. Dengan trafik yang semakin tinggi, maka akses/koneksi internet akan semakin berat/lambat. Sehingga perlu diketahui bagaimana pola trafik internet yang ada selama ini. Pola tersebut berguna untuk dijadikan dasar kebijakan manajemen koneksi internet untuk saat sekarang dan diwaktu yang akan datang. Penelitian ini betujuan untuk melakukan segmentasi pengguna web berdasarkan pola perilaku kunjungan menggunakan metode Genetic K-Means Algorithm. Hasil cluster divalidasi menggunakan metode Silhouette Index. Dengan menggunakan Silhoulette Index dapat diketahui bahwa cluster yang dihasilkan oleh Genetic K-Means Algorithm mengalami peningkatan kualitas sebesar 28,71% lebih baik dibandingkan dengan cluster yang dihasilkan oleh K-Means. Hal ini berarti Genetic K-Means Algorithm bisa mendapatkan cluster yang lebih homogen dan memiliki heterogenitas yang tinggi antar clusternya dibandingkan dengan K-Means.

Kata kunci: Genetic K-Means Algorithm, Silhouette Index, Segmentasi.

Halaman ini sengaja dikosongkan

WEB USERS SEGMENTATION USING GENETIC K-MEANS ALGORITHM

By : Nur Ulfatur Roiha
Student Identity Number : 2214206701
Supervisor(s) : 1. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.
2. Dr. Adhi Dharma Wibawa, ST., MT.

ABSTRACT

Dependance and demand for internet is increasing that causes the increase of internet traffic in daily needs base. The higher traffic affects to internet access connection that will become heavier or slower. Thus, it is necessary to know how the internet traffic pattern occurs every day. The pattern brings an advantage in order to make internet connection management policy for the present and the future time conditions. The research aims is to create web user segmentation based on a web behavior pattern by using Genetic K-Means Algorithm. The cluster result is validated by using Silhouette Index Method that showed that the clusters generated by Genetic K-Means Algorithm has increased for its quality by 28.71% is better than the clusters generated by K-Means. This also has a meaning that Genetic K-Means Algorithm can obtain more homogeneous clusters and have high heterogeneity among inter-clusters compared with K-Means.

Key words: Genetic K-Means Algorithm, Silhouette Index, Segmentation

Halaman ini sengaja dikosongkan

KATA PENGANTAR

Alhamdulillah, segala puji bagi Allah SWT, kita memuji-Nya, meminta pertolongan, pengampunan, serta petunjuk-Nya. Kita berlindung kepada Allah SWT dari kejelekan diri kita dan keburukan amal kita. Semoga shalawat dan salam tetap tercurah pada Rasulullah SAW, keluarga, sahabat beserta siapa saja yang mengikuti petunjuknya.

Sekali lagi segala pujian hanyalah milik Allah SWT, Rabb Semesta Alam yang telah memberikan petunjuk serta karunia-Nya sehingga penulis dapat menyelesaikan tesis berjudul "Segmentasi Pengguna Web menggunakan metode Genetic K-means Algorithm".

Tesis ini disusun guna memenuhi salah satu persyaratan memperoleh gelar Magister Teknik (M.T.) dalam bidang keahlian Telematika-CIO, Jurusan Teknik Elektro, Institut Teknologi Sepuluh Nopember Surabaya. Pada kesempatan ini penulis ingin menyampaikan rasa hormat dan ucapan terima kasih kepada:

1. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc atas arahan, bimbingan dan waktu yang telah dicurahkan kepada penulis selama menjadi dosen pembimbing dan perkuliahan.
2. Dr. Adhi Dharma Wibawa, S.T., M.T. atas arahan, bimbingan dan waktu yang telah dicurahkan kepada penulis selama menjadi dosen pembimbing dan perkuliahan.
3. Prof. Dr. Ir. Mauridhi Hery P, M. Eng, Dr. Ir. Endroyono, DEA, Dr. Surya Sumpeno, S.T., M.Sc, selaku dosen penguji dalam sidang tesis yang telah memberikan masukan dan arahan sehingga memberikan wawasan, pengetahuan dan pemahaman baru untuk penyempurnaan tesis ini.
4. Dr. Ir. Djoko Purwanto, M.Eng. selaku Ketua Jurusan Teknik Elektro, yang memberikan ruang dan kesempatan kepada penulis untuk mengembangkan diri dan berkarya.
5. Prof. Ir. Djauhar Manfaat, M.Sc, Ph.D (Alm) selaku Direktur Pasca Sarjana, yang memberikan ruang dan kesempatan kepada penulis untuk mengembangkan diri dan berkarya di Institut Teknologi Sepuluh Nopember.

6. Semua Dosen Jurusan Teknik Elektro terutama Bidang Telematika-CIO ITS yang telah bersedia membagikan ilmunya kepada penulis, mudah-mudahan bermanfaat dan dapat menjadi amal jariyah.
7. Kementerian Komunikasi dan Informatika yang telah memberikan kesempatan penulis untuk dapat menimba ilmu S2 di jurusan Teknik Elektro, Institut Teknologi Sepuluh Nopember.
8. Dr. Diah Puspito Wulandari, S.T., M.Sc atas arahan, masukan, motivasi dalam menyelesaikan tesis ini.
9. Semua Civitas Akademika Institut Teknologi Sepuluh Nopember, atas semua kebersamaan dan dukungan yang selama ini diberikan kepada penulis selama menimba ilmu.
10. Ayahanda Moh. Hasyim, Ibunda Rukilah, dan Ibu mertua Sartiah serta sudara-saudaraku Dewi Maria Ulfa, Muhammad Zainal Arifin, Muhammad Soffan Syamsudin, Holifah Rusmiwati, Dewi Wulandari, Maulana Malik Ibrahim atas semua dukungan, bantuan dan doa yang tiada terputus selama penulis belajar di S-2.
11. My Beloved Husband, Dedy Riyanto atas semua dukungan, pengertian, motivasi, pengorbanan jiwa dan raga maupun ketabahan serta kesabaran yang luar biasa selama mendampingi penulis menyelesaikan studi S-2. Semoga Allah SWT mencatat segala kebaikan yang dilakukan untuk penulis sebagai amal shaleh.
12. My Beloved Daughter, Fathiyah Putri Riyanto atas semua rasa cinta dan kasih sayang yang dihadirkan untuk penulis. Semoga menjadi anak yang sholeha, berakal cerdas, berilmu lagi beramal serta beriman dan bertakwa kepada Allah SWT.
13. Rekan-rekan S2 Telematika-CIO dan S2 Telematika seangkatan maupun beda angkatan yang selalu memberikan keceriaan, dukungan, motivasi maupun bantuan lainnya.
14. Tim 7 atas segala kebersamaan, motivasi, dukungan dan bantuan lainnya. Suka dan duka telah kita lalui bersama. Tim 7, The Best Friends Forever.

15. Ir. Antiek Sugiharti, M.Si, Hefli Syarifuddin Madjid, SE, M.Si, Emadarta Tri Wijaya ST., MT, yang memberikan ijin bagi penulis untuk menyelesaikan studi S2.
16. Rekan-rekan Bidang Aplikasi dan Telematika, Dinas Komunikasi dan Informatika, Pemerintah Kota Surabaya, Bapak Yudho, Bapak Siswo, Ibu Lina, Ibu Fidya, Ibu Rizka, Bapak Ipoet, Mas Tito, Mas Hadi, Mas Arief, Mbak Aina, Mbak Afi dan semua rekan di Bidang Aplikasi dan Telematika lainnya atas semua dukungan dan doanya, sehingga penulis dapat menyelesaikan studi S-2.
17. Rekan-rekan Bidang Informasi dan Komunikasi Publik, Dinas Komunikasi dan Informatika, Pemerintah Kota Surabaya, Ibu Chosadilia, Bapak Sumar, Bapak Basuki, Ibu Novi, Ibu Nunuk dan rekan-rekan lainnya di Dinas Komunikasi dan Informatika, Pemerintah Kota Surabaya atas semua dukungan yang diberikan kepada penulis.
18. Para sahabatku, Ibu Atik, Mas Rengga, Mbak Eni Yusriani, Mas Pri, Mbak Pat, Mas Irsal, Mbak Rita, Mbak Ike atas semua dukungan, doa yang diberikan kepada penulis.
19. Kepada semua pihak yang telah membantu, mendoakan, memberikan motivasi dan dorongan serta doa yang tidak dapat saya sebutkan satu persatu.

Penulis menyadari bahwa tesis ini masih sangat jauh dari sempurna. Oleh sebab itu, penulis sangat mengharapkan kritik dan saran yang bersifat membangun agar tesis ini menjadi lebih baik. Akhir kata, penulis berharap tesis ini memberikan manfaat terutama untuk pengembangan ilmu pengetahuan dan dapat menjadi amal kebaikan yang dicatat sebagai amal jariah oleh Allah SWT.

Surabaya, Januari 2017

Nur Ulfatur Roiha

Halaman ini sengaja dikosongkan

DAFTAR ISI

PERNYATAAN KEASLIAN TESIS i
LEMBAR PENGESAHAN iii
ABSTRAK v
ABSTRACT vii
KATA PENGANTAR ix
DAFTAR ISI xiii
DAFTAR GAMBAR xx
DAFTAR TABEL xxiii
BAB 1 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 2
1.3 Batasan Masalah 2
1.4 Tujuan Penelitian 2
1.5 Manfaat Penelitian 2
BAB 2 5
2.1 Kajian Pustaka 5
2.2 Penelitian Terdahulu 6
2.3 Clustering 7
2.4 K-Means Clustering 7
2.5 Algoritma Genetika 9
2.5.1. Membangkitkan Populasi Awal 10
2.5.2. Evaluasi Fitness 12
2.5.3. Seleksi 12
2.5.4. Cross Over 14

2.5.5. Mutasi Gen 15
2.5.6. Elitisme 16
2.6 Genetic K-Means Algorithm 16
2.7 Silhouette Index 18
2.8 SARG (Squid Analysis Report Generator) 20
BAB 3 23
3.1 Preprosesing Data 24
3.1.1. Pemilihan Data 24
3.1.2. Penyiapan Database Lokal 24
3.1.3. Menyiapkan Aplikasi Pengolah Database 25
3.1.4. Preprosesing pada URL 26
3.1.4.1. Penentuan parameter 26
3.1.4.2. Kategori Website 27
3.1.5. Pre Processing pada Ukuran Data 29
3.1.6. Normalisasi 29
3.2 Pengklasteran Data 29
3.3 Clustering K-Means dengan Algoritma Genetika 30
3.3.1. Algoritma Genetika 30
3.4 Evaluasi dan Validasi Cluster 36
3.5 Interpretasi Data 36
3.6 Kesimpulan 37
BAB 4 39
4.1 Tahap Praprosesing Data 39
4.1.1. Pemilihan Fitur Data 39
4.1.2. Data Cleaning 43
4.1.3. Transformasi Data 44

4.1.4 Identifikasi URL 46
4.1.5 Normalisasi 50
4.2 Uji Coba 53
4.2.1. Perbandingan Nilai.....54
4.2.2. Perubahan Probabilitas Mutasi 57
4.2.3. Jumlah Pengguna 68
4.2.4. Pengkategorian Akses User 69
BAB 5 73
5.1 Kesimpulan..... 73
5.2 Saran..... 73
DAFTAR PUSTAKA 75
BIODATA PENULIS 77

Halaman ini sengaja dikosongkan

TABLE OF CONTENTS

STATEMENT OF AUTHENTICITY THESIS i
VALIDITY SHEET iii
ABSTRAK v
ABSTRACT vii
PREFACE ix
TABLE OF CONTENTS xiii
LIST OF FIGURE xx
LIST OF TABLE xxiii
CHAPTER 1 INTRODUCTION 1
1.1 Background 1
1.2 Formulation of the Problem 2
1.3 Scope of Problem 2
1.4 Purpose of Research 2
1.5 Benefits of Research 2
CHAPTER 2 LITERATURE REVIEW 5
2.1 Reader Review 5
2.2 Research Accomplished 6
2.3 Clustering 7
2.4 K-Means Clustering 7
2.5 Genetic Algorithm 9
2.5.1. Generating Initial Population 10
2.5.2. Fitness Evaluation 12
2.5.3. Selection 12
2.5.4. Cross Over 14

2.5.5. Gene Mutation 15
2.5.6. Elitism 16
2.6 Genetic K-Means Algorithm 16
2.7 Silhouette Index 18
2.8 SARG (Squid Analysis Report Generator) 20
BAB 3 23
3.1 Data Preprocessing 24
3.1.1. Selection of Data 24
3.1.2. Local Database Setup 24
3.1.3. Preparing Application of Database Processing 25
3.1.4. Preprocesing of URL 26
3.1.4.1. Determining of Parameter 26
3.1.4.2. Website Categorization 27
3.1.5. Pre Processing on Data Size 29
3.1.6. Normalization 29
3.2 Data Clustering 29
3.3 K-Means Clustering with Genetic Algorithms 30
3.3.1. Genetic Algorithm 30
3.4 Evaluation and Validation Cluster 36
3.5 Interpretation of Data 36
3.6 Conclusion 37
BAB 4 39
4.1 Preprocesing of Data 39
4.1.1. Selection of Data Feature 39
4.1.2. Data Cleaning 43
4.1.3. Data Transformation 44

4.1.4 Identifikation of URL 46
4.1.5 Normalization 50
4.2 Testing 53
4.2.1. Comparison of Value 54
4.2.2. Change of Probability Mutation 57
4.2.3. Number of Users 68
4.2.4. Categorization of User Access 69
BAB 5 CONCLUSIONS 73
5.1 Conclusion 73
5.2 Recommendations 73
REFERENCES 75
BIBLIOGRAPHY 77

Halaman ini sengaja dikosongkan

DAFTAR GAMBAR

Gambar 2.1 Siklus Algoritma Genetika oleh David Goldberg 9
Gambar 2.2 Representasi Kromosom 10
Gambar 2.3 Representasi Individu 11
Gambar 2.4 Representasi Kromosom Pembentuk Individu 11
Gambar 2.5 Representasi Gen dan Alele 11
Gambar 2.6 Perbedaan Alele, Gen dan Kromosom 11
Gambar 2.7 Mesin Roulette 13
Gambar 2.8 Proses Cross over 14
Gambar 2.9 Proses Mutasi 15
Gambar 2.10 Proses Eletisme 16
Gambar 2.11 Kekurangan Metode K-means 17
Gambar 2.12 Langkah-langkah Genetic K-Means Algorithm 18
Gambar 2.13 Aplikasi SARG 21
Gambar 3.1 Diagram Penelitian 23
Gambar 3.2 Kategorisasi Website 27
Gambar 3.3 Pengkategorian URL 28
Gambar 3.4 Contoh Ukuran Data 29
Gambar 3.5 Individu yang dibangkitkan 31
Gambar 3.6 Proses Seleksi 34
Gambar 3.7 Proses Cross over 34
Gambar 3.8 Offspring Hasil Cross over 35
Gambar 3.9 Penambahan Individu melalui Proses Cross over 35
Gambar 4.1 Contoh Data Awal 39
Gambar 4.2 Fitur-fitur yang digunakan di tahapan selanjutnya 40
Gambar 4.3 Master URL 40
Gambar 4.4 Proses mengimpor data dari format .csv 41
Gambar 4.5 Data Hasil Impor 42
Gambar 4.6 Struktur Tabel yang telah memiliki Primary Key 42
Gambar 4.7 Tabel yang telah memiliki Primary Key 43

Gambar 4.8 Best Fitness pada Kategori Pemerintahan 57
Gambar 4.9 Nilai Silhouette Index pada Kategori Pemerintahan 58
Gambar 4.10 Best Fitness pada Kategori Email 58
Gambar 4.11 Nilai Silhouette Index pada Kategori Email 59
Gambar 4.12 Best Fitness pada Kategori Email 59
Gambar 4.13 Nilai Silhouette Index pada Kategori Media Sosial 60
Gambar 4.14 Best Fitness pada Kategori Blog/Online Shop 61
Gambar 4.15 Nilai Silhouette Index pada Kategori Blog/Online Shop 61
Gambar 4.16 Best Fitness pada Kategori Blog/Online Shop 62
Gambar 4.17 Nilai Silhouette Index pada Kategori Berita 63
Gambar 4.18 Best Fitness pada Kategori Pendidikan/Iptek 63
Gambar 4.19 Nilai Silhouette Index pada Kategori Pendidikan/Iptek 64
Gambar 4.20 Best Fitness pada Kategori Streaming 65
Gambar 4.21 Nilai Silhouette Index pada Kategori Streaming 66
Gambar 4.22 Best Fitness pada Kategori Pornografi 66
Gambar 4.23 Nilai Silhouette Index pada Kategori Pornografi 67
Gambar 4.24 Jumlah Pengguna 68
Gambar 4.25 User Akses 69

DAFTAR TABEL

Tabel 3.1 Contoh URL yang diakses 26
Tabel 3.2 Kategori URL 26
Tabel 3.3 Contoh Data yang akan diproses 31
Tabel 3.4 Proses Pemberian Fitness..... 32
Tabel 3.5 Jarak Kromosom Terdekat..... 33
Tabel 4.1 Pembersihan Data Uji dengan Query..... 43
Tabel 4.2 Contoh Ukuran Data 44
Tabel 4.3 Contoh Kategorisasi Website..... 46
Tabel 4.4 Transformasi Alamat URL 46
Tabel 4.5 Daftar URL yang diakses oleh Pengguna 7 46
Tabel 4.6 Pengelompokan berdasarkan Pengguna..... 47
Tabel 4.7 Pengelompokan Pengguna Berdasarkan Kategori Berita 48
Tabel 4.8 Pengelompokan berdasarkan kategori Blog / Online Shop 48
Tabel 4.9 Pengelompokan Berdasarkan Kategori Pemerintahan..... 48
Tabel 4.10 Pengelompokan Berdasarkan Kategori Media Sosial..... 49
Tabel 4.11 Pengelompokan Berdasarkan Kategori Pendidikan / Iptek 49
Tabel 4.12 Pengelompokan Berdasarkan Kategori Streaming 50
Tabel 4.13 Pengelompokan diluar Kategori yang telah ditetapkan 50
Tabel 4.14 Normalisasi untuk Kategori Berita 51
Tabel 4.15 Normalisasi untuk Kategori Blog/Online Shop 51
Tabel 4.16 Normalisasi untuk Kategori Pemerintahan 51
Tabel 4.17 Normalisasi untuk Kategori Media Sosial 52
Tabel 4.18 Normalisasi untuk Kategori Streaming..... 52
Tabel 4.19 Normalisasi untuk Kategori Pendidikan/Iptek..... 52
Tabel 4.20 Normalisasi untuk Data diluar Kategori yang ditetapkan..... 53
Tabel 4.21 Perbandingan Nilai Silhouette Index untuk Kategori Pemerintahan .. 54
Tabel 4.22 Perbandingan Nilai Silhouette Index untuk Kategori Email 54
Tabel 4.23 Perbandingan Nilai Silhouette Index untuk Kategori Media Sosial ... 54

Tabel 4.24 Perbandingan Nilai Silhouette Index untuk Kategori Blog/Online Shop 55
Tabel 4.25 Perbandingan Nilai Silhouette Index untuk Kategori Berita 55
Tabel 4.26 Perbandingan Nilai Silhouette Index untuk Kategori Pendidikan/Iptek 55
Tabel 4. 27 Perbandingan Nilai Silhouette Index untuk Kategori Streaming 56
Tabel 4.28 Perbandingan Nilai Silhouette Index diluar Kategori yang telah ditetapkan.....56
Tabel 4.29 Perbandingan Nilai Silhouette Index untuk Kategori Pornografi.....56
Tabel 4.30 Nilai Silhouette Index pada Kategori Pemerintahan 57
Tabel 4.31Nilai Silhouette Index pada Kategori Email.....59
Tabel 4.32 Nilai Silhouette Index pada Kategori Media Sosial 60
Tabel 4.33 Nilai Silhouette Index pada Kategori Blog/Online Shop 61
Tabel 4.34 Nilai Silhouette Index pada Kategori Berita 62
Tabel 4.35 Nilai Silhouette Index pada Kategori Pendidikan/Iptek.....64
Tabel 4.36 Nilai Silhouette Index pada Kategori Streaming.....65
Tabel 4.37 Nilai Silhouette Index pada Kategori Pornografi 67
Tabel 4.38 Jumlah pengguna.....68
Tabel 4.39 Pengkategorian Akses User.....69
Tabel 4.40 Matrik Korespondensi 70

BAB 1
PENDAHULUAN

1.1 Latar Belakang

Ketergantungan terhadap internet semakin meningkat dari waktu ke waktu seiring dengan meningkatnya jumlah pengguna dan kebutuhan terhadap internet. Internet sangat berperan besar hampir diseluruh bidang kehidupan, baik itu di bidang pendidikan, kesehatan, perdagangan, perbankan, perijinan dan berbagai bidang lainnya. Jika kita amati hampir di setiap kantor, perusahaan kecil, menengah, rumah tangga menggunakan layanan jasa internet. Bahkan setiap individu menggunakan gadget yang terhubung langsung dengan internet untuk mengakses situs-situs media sosial seperti facebook, twitter, situs jual beli dan berbagai situs lainnya.

Pemerintah daerah pun tak ketinggalan berlomba-lomba menggunakan aplikasi e-government yang berbasis internet untuk meningkatkan layanan kepada masyarakat maupun untuk kebutuhan internal kantor. Penggunaan aplikasi e-government sendiri bertujuan untuk menjadikan pelayanan publik semakin cepat dan transparan yang tentunya membutuhkan bandwidth internet yang tidak sedikit.

Pemerintah daerah juga menyediakan layanan internet gratis di ruang-ruang publik seperti kantor-kantor perijinan, terminal bus dan angkutan umum, taman-taman kota, sentra PKL (Pedagang Kaki Lima) dan ruang publik lainnya. Seluruh pelayanan ini tentu saja membutuhkan bandwidth internet yang cukup besar. Padahal harga bandwidth relatif cukup mahal dan menyedot anggaran yang cukup banyak.

Namun karena banyaknya pemakaian Internet, aksesnya pun menjadi lambat sehingga diperlukan segmentasi pengunjung web dan perlu pula diketahui pola trafik internet yang ada selama ini. Hal ini diperlukan guna memastikan bahwa internet yang telah disediakan digunakan sesuai dengan tujuann yang telah ditetapkan.

1.2 Rumusan Masalah

Perumusan masalah dalam tesis ini adalah sebagai berikut:

1. Sumber daya jaringan TIK sering disalahgunakan untuk kepentingan diluar kantor / pekerjaan.
2. Aktivitas online pegawai Pemerintah Kota Surabaya tidak terpantau / tidak termonitor secara baik.
3. Belum adanya informasi yang valid tentang pola perilaku pengguna web pada Pemerintah Kota Surabaya.

1.3 Batasan Masalah

Dalam tesis ini, batasan masalah yang dibahas diuraikan sebagai berikut:

1. Data yang digunakan dalam penelitian ini adalah data dari aplikasi SARG yang dikelola oleh Dinas Komunikasi dan Informatika Kota Surabaya sebanyak 95.369 record.
2. Fitur yang digunakan adalah ip address, jumlah koneksi, URL, durasi dan bandwidth yang dibutuhkan untuk mengakses suatu alamat URL.

1.4 Tujuan Penelitian

Tujuan penelitian yang dilakukan adalah mendapatkan pola perilaku pengguna website pada Pemerintah Kota Surabaya menggunakan untuk Genetic K-Means Algorithm.

1.5 Manfaat Penelitian

1. Mendapatkan informasi tentang kecenderungan perilaku pengguna website pada Pemerintah Kota Surabaya dalam mengakses internet.
2. Memudahkan dalam memantau aktivitas pegawai Kota Surabaya dalam mengakses internet.
3. Memastikan bahwa bandwidth yang telah disediakan oleh Dinas Komunikasi dan Informatika Kota Surabaya digunakan sesuai dengan tujuan yang telah ditetapkan.

4. Membantu dalam pembuatan regulasi manajemen bandwidth secara optimal pada Pemerintah Kota Surabaya di masa mendatang.

Halaman ini sengaja dikosongkan

BAB 2
KAJIAN PUSTAKA DAN DASAR TEORI

2.1 Kajian Pustaka

Tan (2006) mendefinisikan data mining sebagai penggalian informasi yang berguna dari gudang data yang besar. Data mining disebut juga pattern recognition merupakan pengolahan data untuk menemukan pola yang tersembunyi dari data tersebut. Hasil dari pengolahan data dengan metode data mining ini dapat digunakan untuk mengambil keputusan di masa depan (Tan, 2006).

Umumnya data mining digunakan untuk data yang berskala besar dan banyak diaplikasikan di berbagai bidang kehidupan baik industri, kesehatan, pendidikan, pedagangan dan masih banyak lainnya.

Data mining merupakan metode pengolahan data berskala besar oleh karena itu data mining ini memiliki peranan penting dalam bidang industri, keuangan, cuaca, ilmu dan teknologi. Secara umum kajian data mining membahas metode-metode seperti, clustering, klasifikasi, regresi, seleksi variabel, dan market basket analisis (Tan,2006).

Data mining merupakan penambangan atau penggalian atau pemilihan atau pengetahuan dari data yang berjumlah banyak. Data mining merupakan proses untuk menganalisa data dari kacamata yang berbeda dan diringkas sehingga dapat menjadi informasi yang bermanfaat. Data mining umumnya digunakan untuk menemukan pengetahuan atau pola yang tersembunyi pada data.

Data mining adalah proses dalam menganalisa maupun meninjau sekumpulan data untuk menemukan pola atau hubungan yang tidak diduga dan meringkas data secara berbeda dengan sebelumnya dan dipahami dan dimanfaatkan oleh pemilik data. Data mining merupakan proses untuk menganalisa data warehouse atau data yang berjumlah besar sehingga membentuk suatu kecenderungan atau pola yang dapat menjadi informasi yang berguna.

Beberapa hal penting terkait dengan data mining sesuai dengan definisi yang disebutkan sebelumnya bahwa:

- Data mining merupakan proses yang dapat berjalan secara otomatis yang terhadap data yang ada.

- Data yang akan diproses merupakan data warehouse atau data yang berjumlah sangat besar.
- Tujuan dari penggunaan data mining adalah untuk mendapatkan kecenderungan, hubungan atau pola yang kemungkinan memberikan indikasi bermanfaat bagi pemilik data.

2.2 Penelitian Terdahulu

Pada bagian ini akan dijelaskan beberapa penelitian terdahulu yang pernah dilakukan. Genetic K-Means Algorithm telah digunakan untuk beberapa penelitian diantaranya adalah penelitian yang dilakukan Bin Lu dan Fangyuan Ju pada tahun 2012 dengan judul "An optimized genetic K-means clustering algorithm" (Lu, et all. 2012). Pada tahun 2014, Ni Luh Gede Pivin Suwirmayanti melakukan penelitian dengan judul "Optimasi Pusat Cluster K-Prototype dengan Algoritma Genetika" (Suwirmayanti, et all. 2014) .

2.3 Clustering

Clustering merupakan suatu metode data mining dan digunakan untuk mencari data kemudian mengelompokkannya berdasarkan similarity (kemiripan karakteristik) antara satu data dengan data yang lain. Clustering sendiri merupakan salah satu metode data mining yang bersifat tanpa bimbingan/arahan (unsupervised). Hal ini artinya tidak ada guru dan tidak ada training/latihan serta tidak memerlukan target atau output. Clustering sendiri ada dua pengelompokan data, yaitu hierarchical clustering dan non-hierarchical clustering (Tahta, 2012).

Hierarchical clustering merupakan suatu pengelompokan data yang diawali dengan mencari dua obyek yang memiliki kemiripan atau kesamaan yang paling dekat. Kemudian dicari obyek yang memiliki kesamaan terdekat yang kedua. Begitu seterusnya sehingga membentuk suatu hirarki. Mulai dari yang mempunyai kemiripan karakteristik terdekat sampai dengan obyek yang paling tidak mirip.

Metode non-hierarchical berbeda dengan metode hierarchical. Jika pada metode hierarchical, hal pertama yang dilakukan adalah mencari kemiripan terdekat. metode non-hierarchical, hal yang pertama dilakukan adalah menentukan jumlah cluster/kelompok yang hendak dibentuk. K-Means clustering termasuk dalam metode non-hierarchical (Tahta, 2012).

2.4 K-Means Clustering

K-Means clustering merupakan salah satu metode yang menerapkan sistem kerja non-hierarchical. Setiap obyek dikelompokkan berdasarkan cluster/kelompok yang telah dibentuk diawal. Setiap obyek yang mempunyai kemiripan dengan anggota yang berada dalam cluster/kelompok yang sama dibandingkan dengan obyek yang diluar cluster/kelompok mereka. Sehingga masing-masing cluster/kelompok memiliki karakteristik yang unik (Agusta, 2007).

Langkah-langkah untuk mengimplementasikan metode K-Means menurut Santosa (Santosa, 2007) dilakukan berdasarkan tahapan-tahapan:

1. Ditentukan jumlah cluster/kelompok k yang diinginkan.

2. Inisialisasi centroid (titik pusat) dari setiap cluster/kelompok. Umumnya inisialisasi centroid (titik pusat) dilakukan dengan membangkitkan angka secara random/acak.
3. Setelah diketahui titik pusat/centroid-nya, maka setiap obyek akan diukur kedekatannya dengan masing-masing titik pusat/centroid. Obyek yang paling dekat dengan titik pusat/centroid, maka akan menentukan obyek tersebut akan menjadi anggota cluster/kelompok yang mana. Perhitungan jarak antara obyek dan pusat cluster/kelompok dapat dilakukan dengan berbagai metode seperti Euclidean distance maupun Manhattan distance. Menurut Yuhefizar, (Yuhefizar, 2014) Perhitungan jarak dapat menggunakan metode Euclidean distance yang dapat dilihat pada persamaan 1:

$$E_{ij} = \sqrt{\sum_{k=1}^{m} (h_{ik} - k_{jk})^2} \qquad (2.1)$$

dengan :

E_{ij} = jarak antara obyek ke-i dan obyek ke-j

m = jumlah variabel

h_{ik} = data dari obyek ke-i pada variabel ke-k

k_{jk} = data dari obyek ke-j pada variabel ke-k

4. Hitung kembali centroid (titik pusat) berdasarkan data yang mengikuti cluster/kelompok masing-masing
5. Ulangi lagi langkah 3 dan 4 hingga tidak ada centroid/kelompok maupun anggota cluster/kelompok yang berubah tempat/posisi.

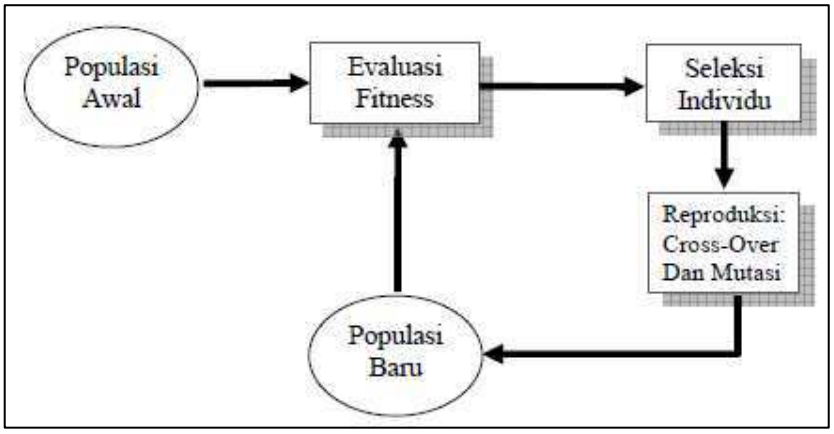
2.5 Algoritma Genetika

Metode algoritma genetika merupakan suatu metode heuristik yang dikembangkan berdasarkan prinsip genetika dan proses alamiah Teori Evolusi Darwin (Zukhri, 2014). Makhluk hidup yang paling kuat dan memiliki ketahanan (fit) paling tinggi yang akan bisa bertahan. Umunya algoritma genetika digunakan untuk menyelesaikan masalah optimasi yang komplek.

Algoritma genetika diciptakan oleh John Holland yang diadapsi dari seleksi alam. Dalam algoritma genetika terdapat populasi yang terdiri atas individu-individu. Individu-individu ini mencerminkan setiap solusi dari permasalahan yang ada. Jika ada 1000 individu maka ada 1000 kemungkinan solusi yang bisa didapatkan.

Setiap individu memiliki nilai ketahanan (fitness) masing-masing. Semakin bagus ketahanan (fitness) yang dimiliki individu maka individu akan semakin mampu bertahan. Sedangkan individu yang memiliki ketahanan (fitness) rendah akan mengalami kepunahan. Begitu pula individu yang berada dalam algoritma genetika, setiap individu yang mencerminkan setiap solusi memiliki nilai ketahanan (fitness) masing-masing. Semakin tinggi nilai fitness-nya maka semakin baik pula solusi yang didapatkan.

David Goldberg pertama kali mengeluarkan siklus Algoritma Genetika yang dapat dilihat pada gambar 2.1:



Gambar 2.1 Siklus Algoritma Genetika oleh David Goldberg

Pada algoritma genetika terdapat populasi awal yang terdiri atas banyak individu. Masing-masing individu akan dilakukan proses evaluasi fitness. Proses ini bertujuan untuk mendapatkan individu yang memiliki ketahanan yang paling

tinggi atau yang paling fit. Individu yang paling fit inilah yang akan dipilih sebagai induk yang selanjutnya akan dilakukan proses kawin silang (cross over) maupun proses mutasi. Dari induk terbaik diharapkan akan didapatkan keturunan (offspring) yang lebih baik dari induknya sehingga akan terbentuk populasi baru yang lebih baik dari populasi sebelumnya.

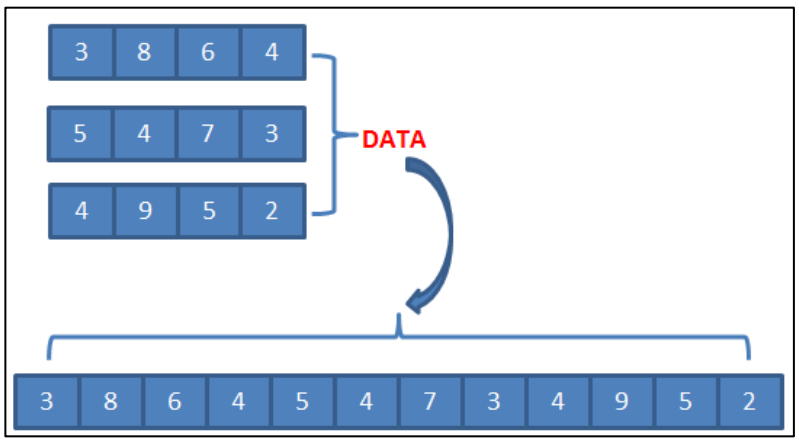
Siklus ini mengalami pengembangan dan diperbaiki oleh Michalewicz. Jika Goldberg menempatkan proses seleksi individu setelah evaluasi fitness dan proses reproduksi setelah seleksi individu, maka Michalewicz mengubah urutan proses Algoritma Genetika yaitu menambahkan proses elitisme setelah dilakukan seleksi individu.

2.5.1. Membangkitkan Populasi Awal

Membangkitkan populasi awal artinya membangkitkan sejumlah individu sebagai anggota dari populasi. Pembangkitan individu dapat dilakukan secara acak/random maupun melalui prosedur tertentu.

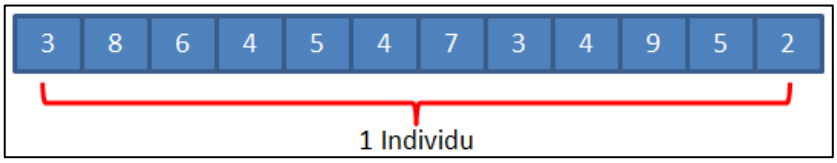
Jumlah individu dalam populasi yang dibangkitkan bergantung pada permasalahan yang ingin diselesaikan. Semakin kompleks permasalahan yang dihadapi maka semakin banyak pula jumlah individu yang dibangkitkan. Semakin banyak jumlah individu maka diharapkan semakin banyak solusi terbaik yang bisa didapatkan.

Umumnya populasi terdiri atas belasan, puluhan, ratusan bahkan ribuan kemungkinan individu atau solusi. Individu terdiri atas kromosom dan kromosom terdiri atas gen. Nilai dari gen disebut dengan alele.



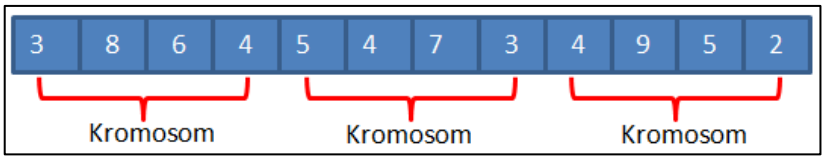
Gambar 2.2 Representasi Kromosom

Satu individu mencerminkan 1 solusi, jika terdapat banyak individu maka akan terdapat banyak kemungkinan solusi permasalahan yang dicari. Representasi individu dapat dilihat pada gambar 2.3.



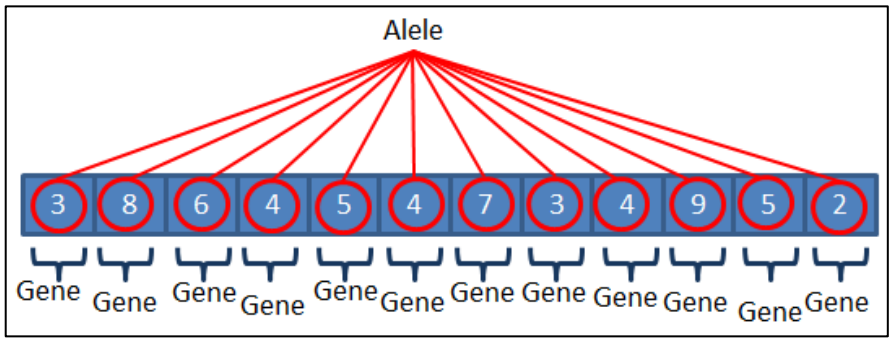
Gambar 2.3 Representasi Individu

Satu individu pada gambar 2.3 terdiri atas 3 kromosom. Representasi kromosom dapat dilihat pada gambar 2.4.



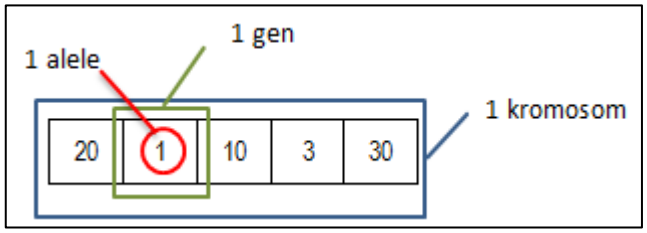
Gambar 2.4 Representasi Kromosom Pembentuk Individu

Satu kromosom terdiri atas banyak gen. Nilai dari gen disebut dengan alele. Nilai alele dibangkitkan secara acak dalam range maksimum dan minimum yang telah ditetapkan sebelumnya. Representasi gen dan alele dapat dilihat pada gambar 2.5.



Gambar 2.5 Representasi Gen dan Alele

Gambaran tentang hubungan antara kromosom, alele dan gen dapat dilihat pada gambar 2.6.



Gambar 2.6 Perbedaan Alele, Gen dan Kromosom

2.5.2. Evaluasi Fitness

Pada Algoritma genetika proses evaluasi fitness dilakukan untuk mendapatkan individu terbaik atau yang mempunyai ketahanan yang paling tinggi (paling fit) dengan melihat nilai fitness dari masing-masing individu/solusi. Individu dengan ketahanan (nilai fitness) terbaik akan dilakukan proses reproduksi yang terdiri atas proses kawing silang dan proses mutasi.

Pada penelitian ini, algoritma genetika digunakan untuk mencari titik pusat cluster/kelompok (centroid) dari masing-masing cluster/kelompok yang mempunyai jarak terpendek dengan masing-masing anggota cluster/kelompok. Menurut Barakbah (Barakbah, 2005), persamaan yang digunakan untuk mencari mencari jarak centroid terpendek dengan anggota cluster-nya dapat dilihat pada persamaan (2.2).

$$J = \sum_{n=1}^{N} \left(min_r \, d(x_n, w_r) \right) \tag{2.2}$$

Sedangkan pencarian nilai fitness dari masing-masing individu yang dibangkitkan dilakukan dengan menggunakan persamaan ().

$$\mathrm{F} = \frac{1}{J} \tag{2.3}$$

dengan:

F = Fungsi Fitness

J = minimum distance

x_n = data ke-n

w_r = centroid ke-r

N = jumlah data

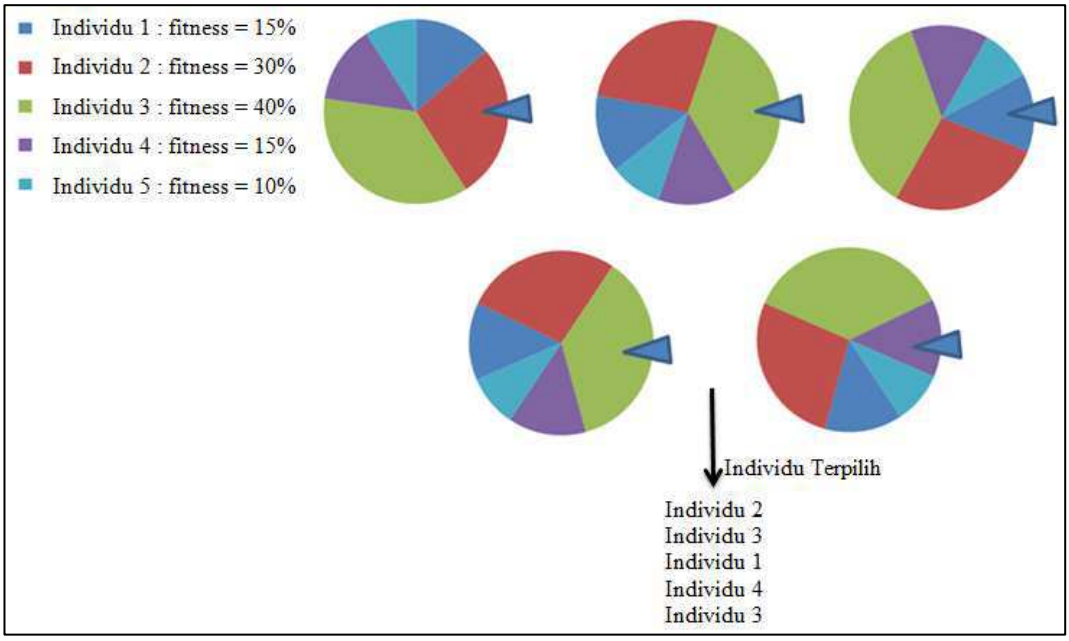
d(y,z) = jarak dari y ke z

2.5.3. Seleksi

Proses seleksi dilakukan bertujuan untuk mendapatkan induk yang baik. Dari induk yang baik inilah diharapkan bisa didapatkan keturunan atau anak (offspring) yang lebih baik dibandingkan dengan induknya namun masih memiliki sebagian sifat induknya. Sebuah individu yang baik dapat dilihat dari nilai fitness-

nya. Semakin tinggi nilai fitness dari sebuah individu maka semakin besar kemungkinannya untuk terpilih.

Proses seleksi umumnya dilakukan menggunakan dua macam teknik, yaitu turnamen dan mesin roullete. Namun dalam penelitian ini proses seleksi dilakukan menggunakan metode mesin roulette. Seleksi menggunakan mesin roulette dapat dilihat pada gambar 2.7.



Gambar 2.7 Mesin Roulette

Setelah melalui proses evaluasi fitness, maka setiap individu memiliki nilai fitness masing-masing. Kemudian dilakukan proses roulette sebanyak jumlah individu. Jika individu berjumlah 4, maka proses roulette dilakukan sebanyak empat kali. Pada gambar 2.7, ditunjukkan bahwa individu 1 memiliki fitness sebesar 0.15, individu 2 sebesar 0.3, individu 3 sebesar 0.4 dan individu 4 sebesar 0.1.

Sebagai contoh, pada proses roulette yang pertama terpilih individu 2, proses roulette kedua terpilih individu 3, proses roulette ketiga terpilih individu 1, proses roulette yang ketiga terpilih individu 4 dan proses roulette yang terakhir terpilih individu 3.

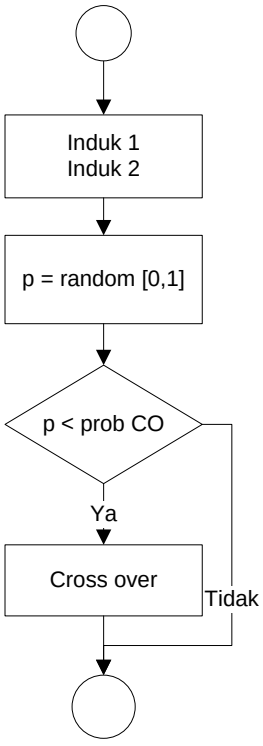
Dengan melakukan proses roulette diharapkan individu yang memiliki fitness terbesar akan sering terpilih.

2.5.4. Cross Over

Cross Over (pindah silang/kawin silang) merupakan operator dalam algoritma genetika yang bertujuan untuk melahirkan individu baru yang memiliki kromosom induknya dan mewarisi sebagian sifat-sifat induknya sebagaimana proses reproduksi yang terjadi dalam kehidupan alam (Zukhri, 2014). Pindah silang membutuhkan dua induk untuk menghasilkan offspring (keturunan) yang baru. Pindah silang dilakukan dengan melakukan pertukaran antar gen secara acak. Proses pindah silang didasarkan pada probabilitas pindah silang yang telah ditentukan sebelumnya.

Langka-langkah dalam proses cross over adalah sebagai berikut:

1. Dilakukan iterasi(i) sebanyak (jumlah individu)/2.
2. Membangkitkan nilai acak (p) antara 0-1.
3. Jika nilai acak (p) < probabilitas cross over (prob CO), maka dilanjutkan ke langkah 4, jika tidak maka kembali ke langkah 2.
4. Merandom 2 angka antara 0 sampai dengan panjang individu (2 angka tersebut adalah batas kiri (bki) dan batas kanan (bka)).
5. Menukar posisi gen dari individu ke-i dan individu ke-(i+1) sepanjang bki sampai dengan bka.

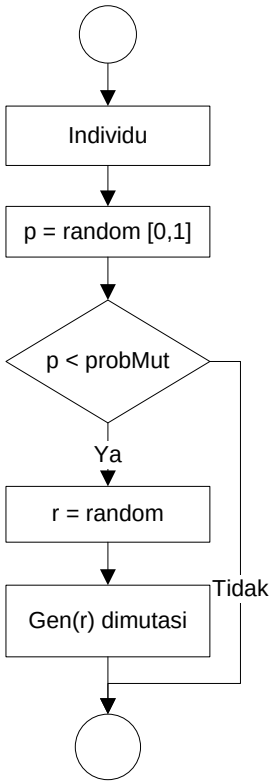


Gambar 2.8 Proses Cross over

2.5.5. Mutasi Gen

Mutasi gen merupakan proses untuk mengubah nilai alele dalam gen dengan nilai inversinya. Jika nilai alele 0 maka diubah menjadi 1, begitu pula sebaliknya. Proses mutasi gen dilakukan juga berdasarkan probabilitas yang telah ditentukan diawal iterasi. Langkah-langkah proses mutasi adalah sebagai berikut:

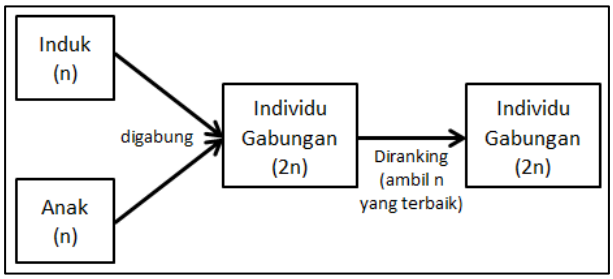
1. Dilakukan iterasi sebanyak jumlah individu.
2. Membangkitkan nilai acak (p) antara 0-1.
3. Jika nilai acak (p) < Probabilitas Mutasi (probMut), maka dilanjutkan ke langkah 4, jika tidak maka dilakukan mutasi. Jika nilai acak (p) > probabilitas maka dilanjutkan ke langkah-4.
4. Dilakukan acak untuk menentukan posisi gen mana yang akan dilakukan mutasi.
5. Dilakukan mutasi terhadap gen yang terpilih pada langkah 4.



Gambar 2.9 Proses Mutasi

2.5.6. Elitisme

Metode elitisme yang dipakai adalah sistem elitisme ranking. Pada metode ini pertambahan individu melalui proses kawin silang maupun mutasi dan individu parents (induk) akan diranking berdasarkan nilai fitness-nya. Kemudian dipilih untuk dijadikan populasi baru. Hal ini yang menyebabkan jumlah individu tetap data tidak bertambah namun individu/solusi memiliki nilai fitness yang lebih bagus. Menurut Barakbah (Barakbah, 2005), proses mutasi sesuai dengan gambar 2.10:



Gambar 2.10 Proses Eletisme

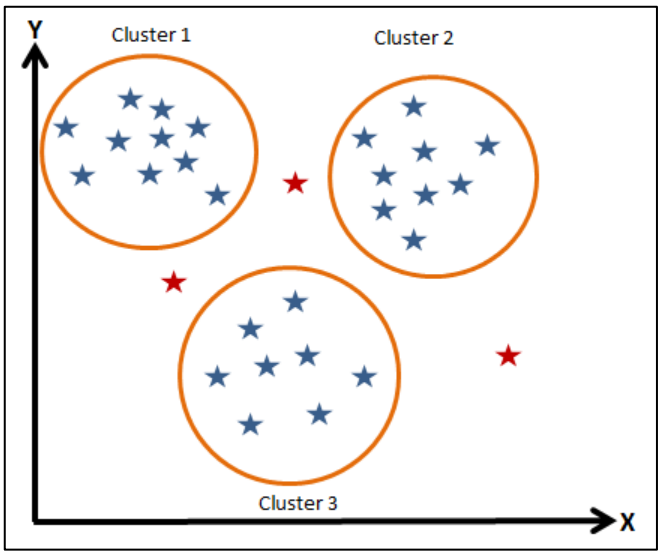
2.6 Genetic K-Means Algorithm

Genetic K-Means Algorithm merupakan metode modifikasi dari metode K-Means. Genetic K-Means Algorithm merupakan gabungan dari metode Algoritma Genetika dan Metode K-Means. Hal ini perlu dilakukan karena titik pusat cluster/centroid yang dibentuk oleh K-Means bersifat acak (random) sehingga cluster/kelompok yang dibentuk tidak optimal. Oleh karena itu digunakan metode algoritma genetika untuk mendapatkan centroid yang optimal yang akan digunakan oleh metode K-Means sehingga K-Means tidak perlu mencari centroid secara random lagi. Dengan digunakan Genetic K-Means Algorithm diharapkan cluster yang terbentuk adalah cluster atau kelompok yang optimal.

Beberapa kekurangan metode K-Means diantaranya adalah :

1. Metode K-Means membangkitkan centroid secara acak/random sehingga centroid yang terbentuk seringkali tidak optimal yang mengakibatkan cluster/kelompok yang terbentuk pun tidak optimal.
2. Memungkinkan suatu cluster/kelompok tidak mempunyai anggota.

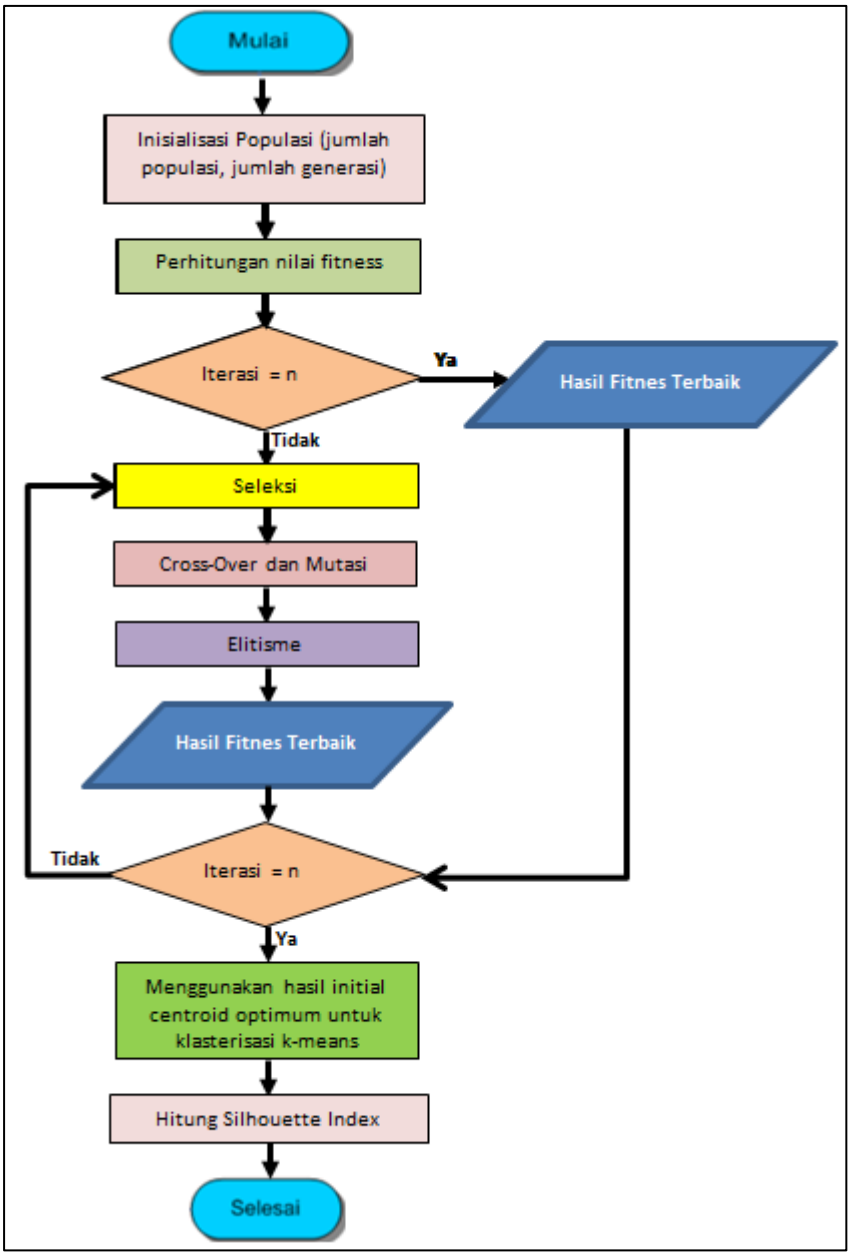
3. Jika terdapat item yang berada di antara dua cluster/kelompok maupun 3 cluster/kelompok (titik-titik kritis) akan menyulitkan penempatan item tersebut termasuk dalam cluster/kelompok tertentu seperti yang terlihat pada gambar 2.11.



Gambar 2.11 Kekurangan Metode K-means

Lu, dkk (2004) mengembangkan algoritma genetika cepat (Fast Genetic K-Means Algorithm) yang digunakan menentukan centroid (pusat cluster/kelompok). Centroid yang dibentuk menggunakan algoritma genetika cepat memiliki performa yang lebih cepat dan menghasilkan cluster yang lebih baik.

Tahapan dalam Metode Genetic K-Means Algorithm dapat dilihat pada gambar 2.12.



Gambar 2.12 Langkah-langkah Genetic K-Means Algorithm

2.7 Silhouette Index

Silhouette index mengacu pada metode penafsiran dan validasi kelompok data. Teknik ini memberikan representasi grafis singkat dari seberapa baik setiap obyek terletak dalam cluster/kelompok. Hal ini pertama kali dijelaskan oleh Peter J. Rousseuw pada tahun 1986 (Santosa, 2007).

Setelah cluster/kelompok terbentuk, maka dilakukan uji validitas menggunakan Silhouette index. Metode ini merupakan metode pengujian untuk:

- Memvalidasi baik sebuah data, cluster tunggal (satu cluster dari sejumlah cluster) atau bahkan keseluruhan cluster.
- Mengukur kualitas/performansi cluster.
- Melihat kualitas dan kekuatan cluster, seberapa baik suatu obyek ditempatkan dalam suatu cluster.

Nilai Silhouette index bervariasi dari -1 dan 1. Nilai Silhouette index mendekati 1 menunjukkan bahwa data tersebut tepat berada pada cluster tersebut. SI bernilai 0 atau mendekati 0 maka posisi data berada pada perbatasan dua cluster. Nilai negatif menandakan jarak rata-rata antar obyek jauh.

Berikut ini merupakan ukuran nilai silhouette index menurut Kaufman dan Rousseuw (Kaufman & P. J. Rousseuw, 1990). Nilai Silhouette index (SI):

- 0.7 < SC <= 1 strong structure
- 0.5 < SC <= 0.7 medium structure
- 0.25 < SC <= 0.5 weak structure
- SC <= 0.25 no structure

Untuk menghitung nilai Silhouette index perlu dari data ke-i digunakan persamaan 2.4:

$$a_i^j = \frac{1}{m_j - 1} \sum_{\substack{r=1 \\ r \neq i}}^{m_j} d(x_i^j, x_r^j), \qquad i = 1,2, \dots, m_j$$

$$b_i^j = \min_{\substack{n=1,\dots,k \\ n \neq j}} \left\{ \frac{1}{m_n} \sum_{\substack{r=1 \\ r \neq i}}^{m_n} d(x_i^j, x_r^n) \right\}, \qquad i = 1,2, \dots, \mathrm{m}_n \tag{2.4}$$

$$SI_i^j = \frac{b_i^j - a_i^j}{\max \left\{ a_i^j, b_i^j \right\}}$$

dengan :

a_i = Rata-rata jarak dari data ke-i terhadap semua data lainnya dalam satu cluster/kelompok

b_i = Rata-rata jarak dari data ke-i terhadap semua data dari cluster/kelompok lain

$d(x_i^r, x_i^r)$ = Jarak data ke-i dengan data ke-r dalam satu cluster/kelompok j

m_j = Jumlah data dalam cluster/kelompok ke-j

SI_i^j = Nilai Silhouette index pada data ke-i

Untuk mendapatkan nilai Silhouette index dari dari sebuah cluster/kelompok dapat dilakukan dengan menghitung nilai rata-rata Silhouette index dalam cluster tersebut dan dapat dilihat pada persamaan 2.5.

$$SI_i^j = \frac{1}{m_j} \sum_{i=1}^{m_j} SI_i^j \quad (2.5)$$

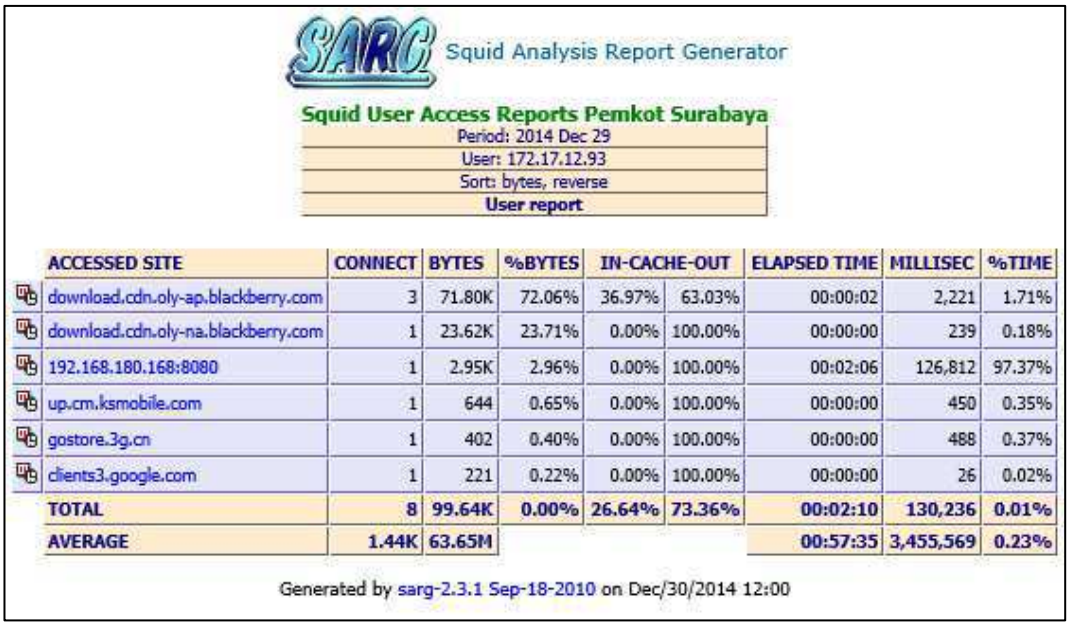
Untuk mendapatkan nilai Silhouette index global dilakukan dengan menghitung rata-rata dari keseluruhan nilai dapat dilihat pada persamaan 2.6.

$$SI = \frac{1}{k} \sum_{j=1}^{k} SI \quad (2.6)$$

Rata-rata Silhouette index dari seluruh data dalam suatu cluster menunjukkan seberapa dekat kemiripan data dalam suatu cluster/kelompok yang juga menunjukkan seberapa tepat data telah dikelompokkan. Semakin dekat Silhouette index kepada 1, maka semakin baik pengelompokkan data Sebaliknya, semakin dekat Silhouette index kepada -1, maka semakin buruk pengelompokkan data.

2.8 SARG (Squid Analysis Report Generator)

SARG (Squid Analysis Report Generator) merupakan tools berbasis open source yang digunakan untuk menganalisa file log squid atau proxy sehingga didapatkan laporan tentang pengguna berupa alamat ip address, situs yang diakses, jumlah bytes yang diakses, waktu yang dibutuhkan untuk mengakses situs tersebut.



Squid Analysis Report Generator

Squid User Access Reports Pemkot Surabaya

Period: 2014 Dec 29
User: 172.17.12.93
Sort: bytes, reverse
User report

ACCESSED SITE	CONNECT	BYTES	%BYTES	IN-CACHE-OUT		ELAPSED TIME	MILLISEC	%TIME
download.cdn.oly-ap.blackberry.com	3	71.80K	72.06%	36.97%	63.03%	00:00:02	2,221	1.71%
download.cdn.oly-na.blackberry.com	1	23.62K	23.71%	0.00%	100.00%	00:00:00	239	0.18%
192.168.180.168:8080	1	2.95K	2.96%	0.00%	100.00%	00:02:06	126,812	97.37%
up.cm.ksmobile.com	1	644	0.65%	0.00%	100.00%	00:00:00	450	0.35%
gostore.3g.cn	1	402	0.40%	0.00%	100.00%	00:00:00	488	0.37%
clients3.google.com	1	221	0.22%	0.00%	100.00%	00:00:00	26	0.02%
TOTAL	**8**	**99.64K**	**0.00%**	**26.64%**	**73.36%**	**00:02:10**	**130,236**	**0.01%**
AVERAGE	**1.44K**	**63.65M**				**00:57:35**	**3,455,569**	**0.23%**

Generated by sarg-2.3.1 Sep-18-2010 on Dec/30/2014 12:00

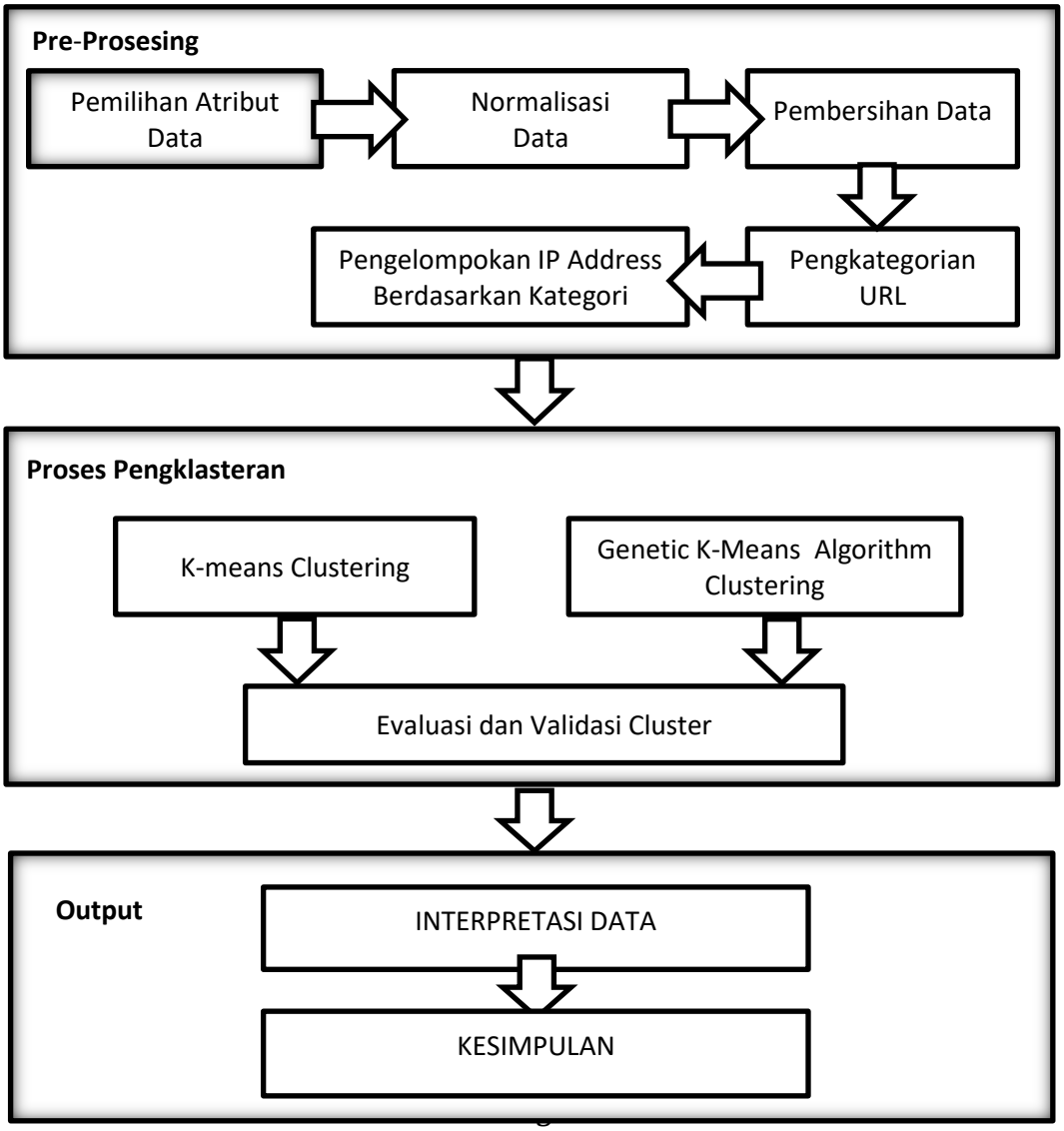
Gambar 2.13 Aplikasi SARG

Proxy merupakan layanan yang dimiliki oleh proxy server. Proxy server bertugas untuk melayani client sedangkan proxy merupakan aplikasi yang menghubungkan antara client dan web server. Proxy bertugas untuk menyimpan cache dari sebuah konten website. Ketika ada salah satu anggota jaringan mengakses sebuah situs, maka tugas proxy menyimpan konten situs tersebut. Jika ada anggota lain dari jaringan yang sama mengakses situs yang sama maka hanya perlu mengakses cache yang telah disimpan dan tidak perlu mengakses internet. Namun jika tidak ada dalam cache maka request tersebut dikirimkan oleh proxy server ke web server.

Halaman ini sengaja dikosongkan

BAB 3
METODE PENELITIAN

Pada tahapan ini akan dijelaskan langkah – langkah metodologi penelitian secara sistematis dan terarah yang akan dijadikan acuan dalam kerangka penelitian yang membahas tentang segmentasi pengguna web menggunakan metode Genetic K-Means Algorithm dan K-Means clustering. Berikut merupakan diagram penelitian yang disajikan pada gambar 3.1 :



Gambar 3.1 Diagram Penelitian

3.1 Preprosesing Data

Pada tahapan ini dijelaskan tentang preprosesing atau pemrosesan awal dari pengolahan data. Preprosesing data merupakan tahapan yang sangat penting dalam penelitian ini. Hal ini disebabkan karena kualitas pengolahan data mining sangat bergantung dari benar tidaknya pada proses preprosesing data. Proses preprocessing juga perlu dilakukan agar data dapat digunakan untuk tahapan berikutnya.

Dalam sebuah penelitian, umumnya data yang telah tersedia tidak semuanya digunakan namun dipilih atribut-atribut atau fitur-fitur data yang menunjang tujuan penelitian tersebut. Data-data yang dipilih tersebut selanjutnya akan dianalisa apakah terdapat nilai yang salah ataupun nilai yang kosong pada satu atau lebih fitur dalam data secara keseluruhan.

3.1.1. Pemilihan Data

Data yang digunakan adalah data yang berasal instansi tempat bekerja penulis yakni Dinas Komunikasi dan Informatika Kota Surabaya. Data diambil dari aplikasi SARG (Squid Analysis Report Generator).

Fitur data yang digunakan pada penelitian ini adalah:

- IP Address.
- Alamat URL.
- Ukuran data yang diakses.
- Lama waktu akses.
- Jumlah Koneksi yang dilakukan.

3.1.2. Penyiapan Database Lokal

Pada penelitian ini diperlukan database lokal yang dapat menyimpan data-data yang akan diteliti. Database yang disediakan berisi banyak tabel. Adapun tabel master yang disiapkan dalam database adalah sebagai berikut:

- Tabel yang digunakan untuk menyimpan data asli yang belum diolah yang berisi data ip address client, url yang diakses, durasi akses dan jumlah bytes yang diakses dan jumlah koneksi.
- Tabel yang berisi tentang kategori website yang diinginkan.
- Tabel yang berisi tentang Top Level Domain (TLD) yang diinputkan untuk masing-masing kategori.
- Tabel Keyword yang digunakan untuk menyimpan masing-masing kategori.

3.1.3. Menyiapkan Aplikasi Pengolah Database

Selain database, dibutuhkan juga aplikasi untuk mengolah data sehingga bisa digunakan untuk melakukan process preprosesing. Menu-menu yang dibutuhkan dalam aplikasi tersebut meliputi:

1. Menu master

   Menu Master berungsi untuk mengatur dan menyimpan data-data yang diperlukan dalam pengkategorian URL yang berisi data master untuk:

   - Kategori; untuk memasukkan kategori website.
   - Domain; untuk memasukkan Top Level Domain (TLD) yang akan dijadikan parameter pada masing-masing kategori.
   - Keyword; untuk memasukkan keyword sebagai parameter dari setiap kategori.
   - Kategorisasi site/URL; untuk mengkategorikan alamat URL berdasarkan domain dan keyword yang dimasukkan.

2. Menu Konversi.

   Menu ini dibutuhkan untuk mengkonversi fitur-fitur yang berbentuk string menjadi numerik sehingga bisa diolah oleh metode K-Means dan Genetic K-Means Algorithm.

3.1.4. Preprosesing pada URL

Dalam melakukan preprosesing pada URL perlu ditentukan parameter dan pengkategorian website. Beberapa contoh data alamat website yang diakses dapat dilihat pada tabel 3.1 dibawah ini:

Tabel 3.1 Contoh URL yang diakses

No	Acessed Site
1	www.yahoo.com
2	eramuslim.com
3	www.detik.com
4	www.bukalapak.com
5	www.olx.com

3.1.4.1. Penentuan parameter

Dalam proses analisis akses internet, diperlukan parameter-parameter yang berkaitan dengan proses kategorisasi website. Dalam proses kategorisasi website, parameter yang perlu diperhatikan adalah:

- Kategori Website, adapun kategori yang digunakan pada penelitian ini dapat dilihat pada tabel 3.2.

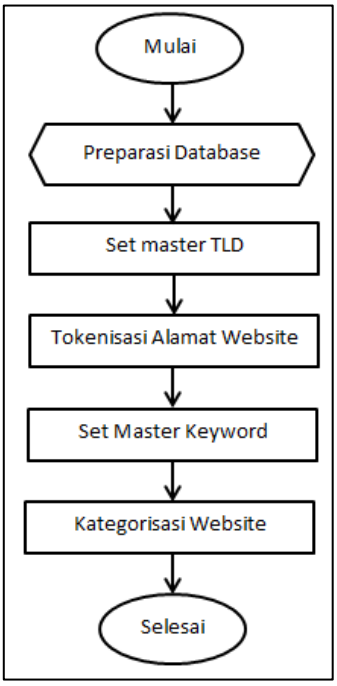
Tabel 3.2 Kategori URL

No	Kategori	URL
1.	Pemerintahan	.go, .gov, .go.id, undang, pemerintah, perundangan, peraturan, kementerian
2.	Pendidikan atau Iptek	.ac. id, .edu, .sch.ac.id, sekolah, pendidikan, universitas, perguruan tinggi
3.	Email	gmail, ymail, mail, pop3, smtp
4.	Blog atau Online Shop	blogspot, wordpress, .blog, blogger
5.	Streaming	.youtube, , skype, chat, messenger, video
6.	Media Sosial	facebook, twitter, instagram, kaskus, twitter,
7.	Berita	news, .detik, kompas, jawapos, liputan
8.	Pornografi	Porn, porno
9.	Lain-lain	.com, .co.id

- Top Level Domain (TLD) dari alamat URL.
- Keyword yang dimasukkan untuk masing-masing kategori.
- Kompleksitas alamat URL yang akan dikategorikan.

3.1.4.2. Kategori Website

Menurut yusriani (Yusriani, 2014) proses kategorisasi website secara garis besar terdiri atas beberapa tahap yang bisa dilihat pada gambar 3.2.



Gambar 3.2 Kategorisasi Website

Proses kategorisasi website dilakukan dengan tahap:

1. Pengecekan TLD dari url yang diakses oleh client.

   Tabel master domain berfungsi untuk menyimpan domain dari alamat url yang diakses oleh client. Setiap domain mengacu pada kategori tertentu. Sebagai contoh domain .go atau .go.id atau .gov termasuk dalam kategori pemerintahan.

2. Tokenizer

   Tokenizer merupakan sebuah tool yang digunakan untuk memisahkan setiap kata yang ada dalam alamat url yang diakses oleh client. Sebagai contoh jika client mengakses alamat url www.sekarfajartimur.blogspot.com maka url tersebut secara otomatis terpisah dalam 4 kata yaitu www, sekarfajartimur, blogspot dan com.

3. Memasukkan token ke dalam master keyword.

   Token-token yang dihasilkan dari proses tokenizer disimpan dalam master keyword, sedangkan setiap keyword mengacu kepada setiap kategori yang

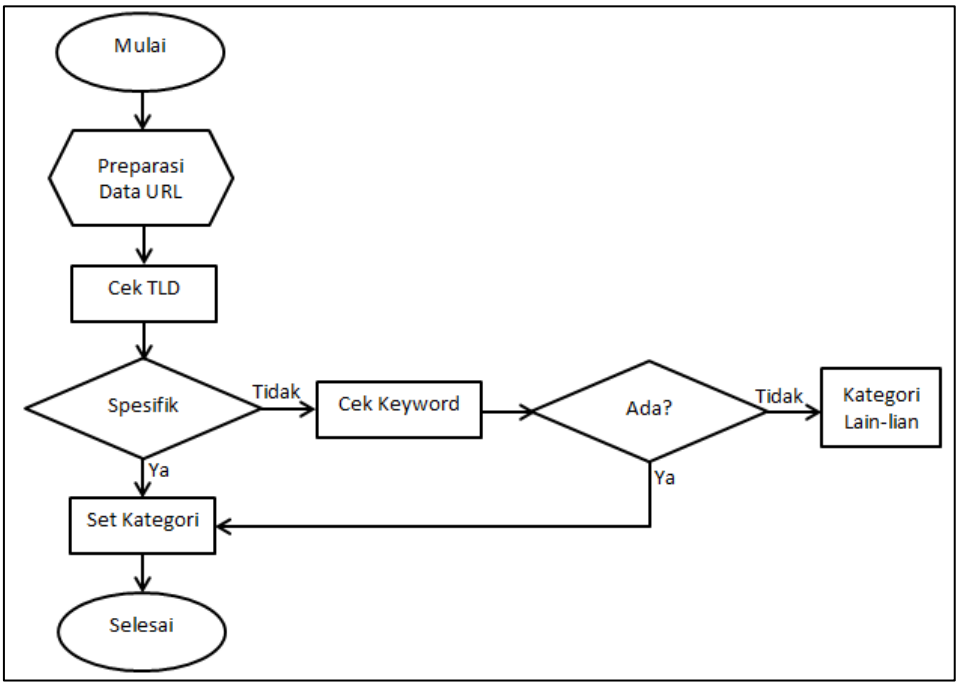
telah ditetapkan. Proses penentuan keyword dilakukan melalui menu pada aplikasi.

4. Membuat master kategori yang dibutuhkan untuk kategorisasi website berdasarkan URL yang terekam database.

   Dalam penelitian ini terdapat 9 kategori, yaitu kategori pemerintahan, pendidikan/iptek, email, blog/online shop, streaming, media sosial, berita, dan pornografi. Sedangkan jika keyword yang ada tidak adalam 8 kategori sebelumnya, maka keyword yang baru dimasukkan dalam kategori lain-lain. Hasil dari pengkategorian disimpan dalam tabel keyword.

5. Mengkategorikan URL yang ada pada tabel sites

   Data site diberikan keterangan yang berupa id_kategori yang menunjukkan kategori dari site tersebut. Kategori yang ada adalah sesuai dengan master kategori yang sudah diinputkan, dengan menggunakan acuan pada TLD dan keyword yang ada pada alamat URL tersebut. URL/site yang sudah dikategorikan, dimasukkan dalam tabel kategori_sites untuk memudahkan pengambilan dan pengolahan data



Gambar 3.3 Pengkategorian URL

3.1.5. Pre Processing pada Ukuran Data

Ukuran data yang diakses oleh client bervariasi ada yang berukuran bytes, kilo bytes maupun mega bytes. Untuk itu perlu distandarkan menjadi bytes. Contoh ukuran dapat dilihat pada Gambar 3.4:

www.intijayapets.com	1	657.21K
jualbeliburung.com	26	434.55K
rma-api.gravity.com	96	397.21K
www.ibo.co.id	24	295.60K

Gambar 3.4 Contoh Ukuran Data

3.1.6. Normalisasi

Normalisasi merupakan proses pengskalaan pada fitur data sehingga data berubah menjadi range tertentu (Barakbah, 2005). Pada penelitian ini, digunakan metode normalisasi min-max. Metode min-max merupakan metode normalisasi dengan cara mentransformasi secara linier terhadap data asli. Persamaan yang digunakan dalam metode min-max adalah:

$$\text{data\_baru} = \frac{(\text{data\_lama\_minimal}) * (\text{data\_baru\_maksimal} - \text{data\_baru\_minimal})}{(\text{skala\_maksimal} - \text{skala\_minimal}) + \text{data\_baru\_minimal}} \quad (3.1)$$

Metode min-max mempunyai kelebihan dibandingkan dengan metode normalisasi lainnya yaitu adanya keseimbangan nilai perbandingan antara data sebelum dilakukan normalisasi dan sesudah normalisasi.

3.2 Pengklasteran Data

Tahapan pengklasteran data dilakukan setelah proses pembersihan dan normalisasi data. Pengklasteran ini dilakukan menggunakan 2 metode, yakni metode K-Means dan metode Genetic K-Means Algorithm. Sedangkan metode Genetic K-Means Algorithm melalui 2 tahap yaitu melalui metode Algoritma Genetika dan K-Means.

Sedangkan tujuan dilakukan pengklasteran menggunakan metode Genetic K-Means Algorithm ialah adanya kelemahan algoritma K-means yang bersifat local optima yaitu hasil yang didapat K-means terkadang baik terkadang jelek, hal

tersebut terjadi karena tidak ada perhitungan pasti untuk membangkitkan pusat centroid awal yang dilakukan algoritma K-means, karena selama ini untuk menentukan pusat centroid awal dengan cara random sehingga jika centroid yang digunakan tepat maka hasil yang didapatkan bagus sebaliknya jika centroid yang digunakan kurang tepat maka hasil yang didapatkan kurang bagus, bagus atau tidak maksudnya adalah jarak kemiripan antar anggota cluster kecil dan jarak antar cluster yang berbeda tinggi. Maka dengan dilakukan pengklasteran dengan Genetic K-Means Algorithm adalah untuk mendapatkan nilai pusat centroid awal yang digunakan untuk algoritma K-means.

3.3 Clustering K-Means dengan Algoritma Genetika

Data yang diproses melalui kegiatan preprosesing menghasilkan data yang berbentuk numerik. Data ini kemudian diproses menggunakan metode Genetic K-Means Algorithm dan K-Means.

3.3.1. Algoritma Genetika

Permasalahan yang hendak diselesaikan dalam penelitian ini adalah mencari centroid yang paling optimal dengan menggunakan algoritma genetika. Centroid yang paling optimal tersebut akan digunakan sebagai centroid dalam metode K-Means sehingga K-Means tidak lagi perlu mencari centroid secara acak.

3.3.1.1. Pembangkitan Individu

Proses algoritma genetika diawali dengan membangkitkan sejumlah individu yang berisi kemungkinan solusi dari permasalahan yang hendak dipecahkan. Semakin banyak individu yang dibangkitkan maka semakin banyak pula solusi yang bisa didapatkan.

Cluster yang dibentuk sejumlah 3 cluster, maka dari itu individu atau solusi yang mungkin adalah berupa centroid dari 3 cluster yang dibentuk. Masing-masing centroid memiliki koordinat tertentu. Jika terdapat 4 fitur maka masing-masing centroid memiliki 4 titik koordinat atau sumbu bidang. Titik koordinat ini disebut sebagai gen. Nilai dari titik koordinat atau gen disebut dengan alele.

Setiap individu mengandung 3 kromosom, dan setiap kromosom terdiri atas 4 gen yang didalamnya terdapat 4 alele. Maka panjang dari individu adalah hasil perkalian dari jumlah fitur dengan jumlah cluster yang ingin dibentuk. Contoh individu yang dibangkitkan dapat dilihat pada Gambar 3.5.

	Kromosom 1						Kromosom 2						Kromosom 3					
	kategori	size	durasi	bulan	hari	konek	kategori	size	durasi	bulan	hari	konek	kategori	size	durasi	bulan	hari	konek
Individu 1	38.13	1.00	1.00	1.00	50.50	1.00	25.75	1.73	1.00	1.00	17.50	5.88	87.63	1.08	1.04	100.00	100.00	17.83
Individu 2	87.63	1.05	1.00	1.06	50.50	9.76	75.25	2.04	1.83	100.00	67.00	50.60	38.13	1.00	1.00	100.00	100.00	2.10
Individu 3	75.25	2.85	1.00	1.00	34.00	26.20	100.00	1.00	1.01	100.00	100.00	4.09	62.88	3.30	2.99	100.00	100.00	1.30
Individu 4	38.13	1.00	1.00	1.00	17.50	1.00	87.63	1.01	1.00	1.01	83.50	1.40	75.25	1.00	1.02	100.00	100.00	1.50

Gambar 3.5 Individu yang dibangkitkan

3.3.1.2. Perhitungan Nilai *Fitness*

Solusi atau centroid yang paling maksimal diharapkan akan berada dintara individu yang dibangkitkan. Untuk itulah diperlukan pencarian jarak terdekat antara data dengan individu yang telah dibangkitkan.

Tabel 3.3 Contoh Data yang akan diproses

	Kategori	Size	Durasi	Bulan	Hari	Konek
Data 1	62.88	76.06	2.94	100.00	1.00	25.10
Data 2	87.62	1.01	1.00	100.00	17.50	4.29
Data 3	75.25	1.00	1.00	1.00	34.00	1.00
Data 4	50.50	1.01	1.00	1.00	50.50	2.00

Proses perhitungan dilakukan dengan cara menghitung jarak euclidean distance (mencari jarak kromosom terdekat) antara setiap kromosom individu yang berada dalam gambar 3.5 dengan setiap data yang berada di tabel 3.12.

1. Perhitungan Jarak antara Data dengan Individu yang dibangkitkan

Perhitungan jarak antara data dengan setiap individu dilakukan dengan menghitung jarak setiap data dengan setiap kromosom yang dimiliki individu menggunakan persamaan:

$$M(v,k,j) = \sqrt{\sum_{i=1}^{n} \left(y(v,k,\mathrm{j}) - x(i,j)\right)^2} \qquad (3.2)$$

dengan:

N = Jumlah data

V = Individu ke-v

k = Kromosom atau cluster ke-k

i = Data ke-i

j = Fitur ke-j

Hasil perhitungan jarak antara setiap data dengan setiap individu dapat dilihat pada tabel 3.4.

Tabel 3.4 Proses Pemberian Fitness

		Kromosom ke-1	Kromosom ke-2	Kromosom ke-3	Nilai Kromosom Terkecil
Individu ke-1	Data ke-1	138.14	179.93	126.85	126.85
	Data ke-2	93.97	95.72	128.76	128.76
	Data ke-3	158.49	127.34	125.15	127.34
	Data ke-4	138.14	161.37	127.08	127.08
					510.03
Individu ke-2	Data ke-1	144.41	116.76	83.60	83.60
	Data ke-2	135.69	68.92	96.23	96.23
	Data ke-3	134.95	83.42	86.23	83.42
	Data ke-4	140.58	119.01	83.47	83.47
					346.73
Individu ke-3	Data ke-1	84.64	52.41	120.80	120.80
	Data ke-2	77.56	115.55	124.64	124.64
	Data ke-3	78.43	121.57	119.66	121.57
	Data ke-4	84.64	51.03	118.98	118.98
					486.00

		Kromosom ke-1	Kromosom ke-2	Kromosom ke-3	Nilai Kromosom Terkecil
Individu ke-4	Data ke-1	51.02	30.15	105.73	105.73
	Data ke-2	62.50	131.27	99.77	131.27
	Data ke-3	48.37	138.05	99.82	138.05
	Data ke-4	20.62	90.47	102.05	102.05
					477.10

2. Perbandingan

Setelah diketahui masing-masing jarak setiap data dengan setiap kromosom yang dimiliki oleh individu dengan maka dilakukan proses pemilihan nilai kromosom terkecil dan dijumlahkan dengan nilai kromosom terkecil lainnya.

Tabel 3.5 Jarak Kromosom Terdekat

No	Individu 1	Individu 2	Individu 3	Individu 4
1	126.85	83.60	120.80	105.73
2	128.76	96.23	124.64	131.27
3	127.34	83.42	121.57	138.05
4	127.08	83.47	118.98	102.05
Jumlah	**510.03**	**346.73**	**486.00**	**477.10**

Nilai fitness masing-masing individu didapatkan dengan menggunakan persamaan :

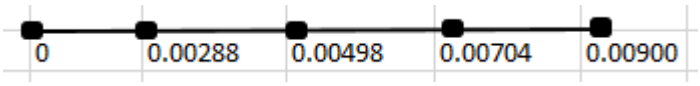
- Nilai fitness individu 1 = 1/510.03 = 0,00196
- Nilai fitness individu 2 = 1/346.73 = 0,00288
- Nilai fitness individu 3 = 1/486.00 = 0,00206
- Nilai fitness individu 4 = 1/477.10 = 0,0021

3.3.1.3. Seleksi Individu

Setelah diketahui nilai fitness dari masing-masing individu, maka dilakukan pemilihan individu untuk dilakukan cross over dan mutasi. Pemilihan individu dilakukan dengan menggunakan sistem roulette dengan tahapan sebagai berikut:

1. Penjumlahan semua fitness untuk dijadikan batasan range untuk proses roulette.

Ranking	Individu	Jarak	Nilai Fitness	Jumlah
1	2	346.73	0.00288	0.00498
2	4	477.10	0.00210	0.00704
3	3	486.00	0.00206	0.00900
4	1	510.03	0.00196	0.00498

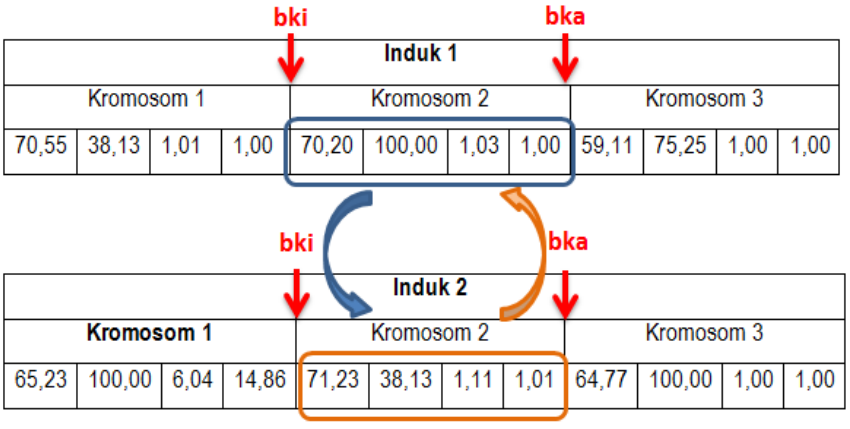


Gambar 3.6 Proses Seleksi

2. Dilakukan roulete sebanyak jumlah individu. Sebagai contoh roulette yang pertama didapatkan angka 0,002 sehingga yang terpilih adalah individu 2 karena 0,021 masuk dalam range individu 2.

3.3.1.4. *Cross over*

Operator Cross over atau kawin silang dilakukan dengan harapan bisa didapatkan keturunan yang lebih baik dibandingkan individu yang ada saat ini. Setelah didapatkan nilai fitness dari masing-masing individu, selanjutnya dilakukan cross over atau kawin silang diantara 2 individu yang telah dibangkitkan. Kawin silang antar individu dilakukan Bagan proses cross over dapat dilihat pada bagan dibawah ini:



Gambar 3.7 Proses Cross over

Hasil dari perkawinan dari 2 induk dapat dihasilkan 2 keturunan (offspring) yang dapat dilihat pada Gambar 3.8.

Anak 1											
Kromosom 1				Kromosom 2				Kromosom 3			
70,55	38,13	1,01	1,00	71,23	38,13	1,11	1,01	59,11	75,25	1,00	1,00

				Anak 2							
Kromosom 1				Kromosom 2				Kromosom 3			
65,23	100,00	6,04	14,86	70,20	100,00	1,03	1,00	64,77	100,00	1,00	1,00

Gambar 3.8 Offspring Hasil Cross over

Melalui proses cross over, individu yang semula berjumlah 2, mengalami penambahan menjadi 4 individu seperti yang dapat dilihat pada gambar 3.9.

Induk 1											
Kromosom 1				Kromosom 2				Kromosom 3			
70,55	38,13	1,01	1,00	70,20	100,00	1,03	1,00	59,11	75,25	1,00	1,00

				Induk 2							
Kromosom 1				Kromosom 2				Kromosom 3			
65,23	100,00	6,04	14,86	71,23	38,13	1,11	1,01	64,77	100,00	1,00	1,00

Anak 1											
Kromosom 1				Kromosom 2				Kromosom 3			
70,55	38,13	1,01	1,00	71,23	38,13	1,11	1,01	59,11	75,25	1,00	1,00

				Anak 2							
Kromosom 1				Kromosom 2				Kromosom 3			
65,23	100,00	6,04	14,86	70,20	100,00	1,03	1,00	64,77	100,00	1,00	1,00

Gambar 3.9 Penambahan Individu melalui Proses Cross over

3.3.1.5. Mutasi

Proses mutasi dari masing-masing individu dilakukan dengan mengubah nilai gen individu menjadi bilangan biner. Kemudian merubah nilai gen menjadi nilai inversinya. Proses mutasi membutuhkan 2 induk, jika terdapat perbedaan nilai

diantara 2 induk, maka dapat dilakukan mutasi sesuai dengan probabilitas mutasi yang telah ditetapkan.

3.3.1.6. Elitisme

Melalui proses cross over dan mutasi, jumlah individu mengalami pertambahan yang cukup banyak. Namun jumlahnya akan dikurangi sesuai dengan penetapan individu di awal proses pembangkitan individu.

Keseluruhan individu tersebut akan melalui proses fitness kembali dan diranking berdasarkan nilai finess yang didapatkan. Sebagai contoh jika diawal proses telah ditetapkan sebanyak n individu. Setelah melalui proses cross over dan mutasi individu bertambah menjadi 2n. Ketika proses elitisme, maka setelah diranking hanya akan diambil sebanyak n individu yang memiliki fitness terbaik, sehingga dalam setiap iterasi jumlah individu tetap namun memiliki nilai fitness yang semakin bagus.

3.4 Evaluasi dan Validasi Cluster

Pada proses evaluasi dilakukan setelah dilakukan proses pengklasteran atau pada saat didalam proses clustering itu sendiri, proses ini penting karena evaluasi cluster ini untuk menguji data set yang digunakan untuk melihat kualitas dan kekuatan cluster, seberapa baik suatu objek ditempatkan dalam suatu cluster. Jika suatu cluster mempunyai kualitas baik maka tingkat homogenitas cluster tinggi, dalam penentuan pola dan analisa cluster akan semakin mudah. Sedangkan evaluasi dan validari cluster pada penelitian ini adalah menggunakan metode Silhouette index (SI).

3.5 Interpretasi Data

Pada bagian ini merupakan proses intepretasi data yaitu setelah dilakukan pengklasteran dan optimalisasi jumlah cluter maka didapat hasil pengelompokan yang optimal langkah berikutnya ialah mendeskripsikan pola hasil dari pengelompokan tersebut.

3.6 Kesimpulan

Pada tahap ini merupakan proses akhir yaitu menyimpulkan dari hasil yang sudah didapat, kemudian manfaat apa yang didapatkan dari hasil penelitian ini ialah untuk menentukan berbagai kebijakan untuk sebagai bahan pertimbangan.

Halaman ini sengaja dikosongkan

BAB 4
HASIL PENELITIAN DAN PEMBAHASAN

4.1 Tahap Praprosesing Data

Data untuk penelitian ini didapatkan dari Dinas Komunikasi dan Informatika Kota Surabaya sejumlah 95.369 record. Data ini merupakan data yang diambil dari Aplikasi SARG (Squid Analysis Report Generator) yang berisi tentang data trafik internet yang ada di Pemerintah Kota Surabaya.

Tahapan awal yang dilakukan pada penelitian ini adalah pemilihan fitur data, transformasi data dan pembersihan data. Dari tahapan awal ini diharapkan akan mendapatkan data awal yang dapat diproses pada tahapan selanjutnya.

4.1.1. Pemilihan Fitur Data

Data yang tersimpan didapatkan dari aplikasi SARG (Squid Analysis Report Generator) adalah data yang berbentuk text. Data tersebut kemudian ditransformasi dalam format excel. Kemudian dilakukan pemilihan fitur yang akan digunakan berdasarkan kolom-kolom data yang diperlukan. Data awal dapat dilihat pada Gambar 4.1.

DATE	IP	ACCESSED SITE	CONNECT	BYTES	ELAPSED TIME	MILLISEC
29/12/2014	172.17.21.205	dlcdnet.asus.com	243	3.48G	0:40:45	2,445,714
29/12/2014	172.17.21.205	s.kaskus.id	234	3.59M	0:00:07	7,879
29/12/2014	172.17.21.205	images.detik.com	76	1.84M	0:00:01	1,13
29/12/2014	172.17.21.205	cdn.kaskus.com	19	1.74M	0:00:07	7,532
29/12/2014	172.17.21.205	connect.facebook.net	12	632.00K	0:00:00	232
29/12/2014	172.17.21.205	newopenx.detik.com	97	557.22K	0:00:10	10,027
29/12/2014	172.17.21.205	detik.net.id	64	509.62K	0:00:02	2,204
29/12/2014	172.17.21.205	4.bp.blogspot.com	11	423.54K	0:00:02	2,073
29/12/2014	172.17.21.205	cdn.maskool.in	5	316.76K	0:00:01	1,958
29/12/2014	172.17.21.205	www.google-analytics.com	67	278.86K	0:00:05	5,025
29/12/2014	172.17.21.205	www.tupperware.co.id	7	225.46K	0:00:00	368
29/12/2014	172.17.21.205	www.detik.com	31	195.10K	0:00:05	5,965
29/12/2014	172.17.21.205	www.kaskus.co.id	11	194.53K	0:00:06	6,864

Gambar 4.1 Contoh Data Awal

Dari data awal yang berbentuk excel, dilakukan pemilihan kolom yang akan dijadikan fitur untuk tahapan selanjutnya. Data yang dijadikan fitur adalah data IP Address, Accessed Site, Bytes, Millisec, Connect seperti yang terlihat pada gambar 4.2.

IP	URL	Bytes	Millisec	Connect
172.17.21.205	dlcdnet.asus.com	3.48G	2,445,714	243
172.17.21.205	s.kaskus.id	3.59M	7,879	234
172.17.21.205	images.detik.com	1.84M	1,13	76
172.17.21.205	cdn.kaskus.com	1.74M	7,532	19
172.17.21.205	connect.facebook.net	632.00K	232	12
172.17.21.205	newopenx.detik.com	557.22K	10,027	97
172.17.21.205	detik.net.id	509.62K	2,204	64
172.17.21.205	4.bp.blogspot.com	423.54K	2,073	11
172.17.21.205	cdn.maskool.in	316.76K	1,958	5
172.17.21.205	www.google-analytics.com	278.86K	5,025	67
172.17.21.205	www.tupperware.co.id	225.46K	368	7
172.17.21.205	www.detik.com	195.10K	5,965	31
172.17.21.205	www.kaskus.co.id	194.53K	6,864	11

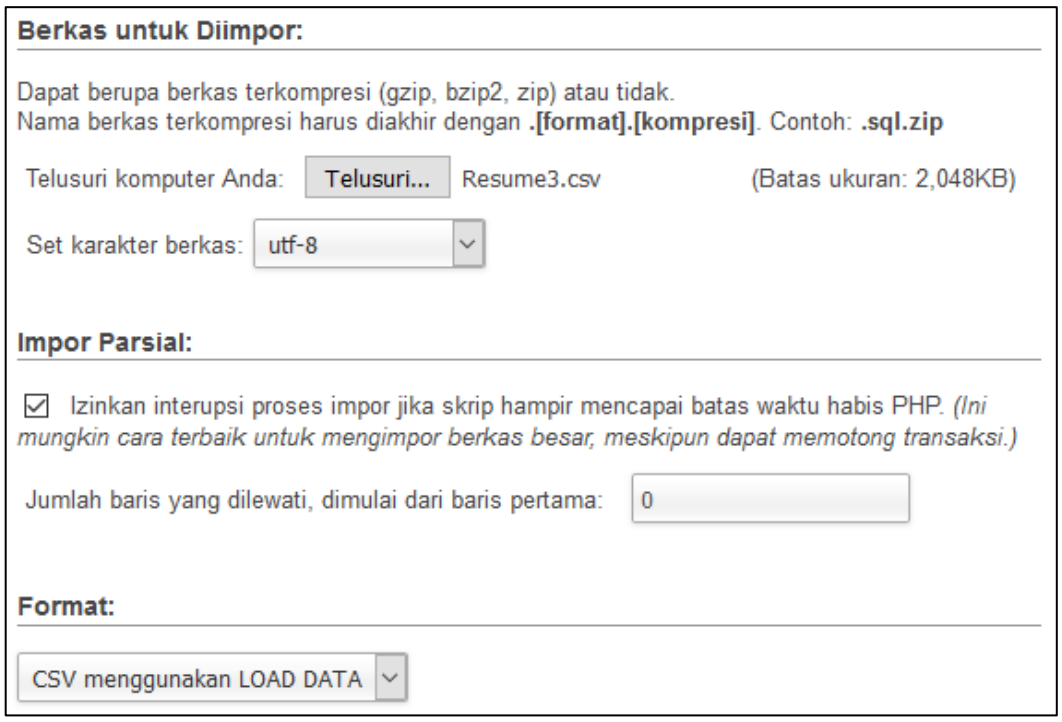
Gambar 4.2 Fitur-fitur yang digunakan di tahapan selanjutnya

Fitur-fitur yang digunakan disimpan dalam format .csv dan disimpan dalam database. Database Management System (DBMS) yang digunakan adalah aplikasi MySQL. Data dalam format .csv diimpor dalam database dan tabel yang telah dibuat sebelumnya seperti yang terlihat pada gambar 4.3.

#	Nama	Jenis	Penyortiran	Atribut	Kosong	Bawaan
1	**bulan**	int(11)			Tidak	*Tidak ada*
2	**tanggal**	int(11)			Tidak	*Tidak ada*
3	**ip**	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*
4	**site**	varchar(500)	latin1_swedish_ci		Tidak	*Tidak ada*
5	**bytes**	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*
6	**millisec**	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*
7	**konek**	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*

Gambar 4.3 Master URL

Ip digunakan untuk menyimpan ip address pengguna/user, sedangkan site digunakan untuk menyimpan alamat URL dari website yang diakses oleh pengguna, bytes digunakan untuk menyimpan ukuran file yang diakses, millisec digunakan untuk menyimpan durasi yang dibutuhkan untuk mengakses suatu website dan konek digunakan untuk menyimpan jumlah koneksi yang diperlukan dalam mengakses website.



Gambar 4.4 Proses mengimpor data dari format .csv

Hasil impor data dari format .csv ke tabel yang sudah disediakan, dapat dilihat pada gambar 4.5.

bulan	tanggal	ip	site	bytes	millisec	konek
12	1	172.17.21.205	dlcdnet.asus.com	3.48G	2,445,714	243
12	1	172.17.21.205	s.kaskus.id	3.59M	7,879	234
12	1	172.17.21.205	images.detik.com	1.84M	1,13	76
12	1	172.17.21.205	cdn.kaskus.com	1.74M	7,532	19
12	1	172.17.21.205	connect.facebook.net	632.00K	232	12
12	1	172.17.21.205	newopenx.detik.com	557.22K	10,027	97
12	1	172.17.21.205	detik.net.id	509.62K	2,204	64
12	1	172.17.21.205	4.bp.blogspot.com	423.54K	2,073	11
12	1	172.17.21.205	cdn.maskool.in	316.76K	1,958	5
12	1	172.17.21.205	www.google-analytics.com	278.86K	5,025	67
12	1	172.17.21.205	www.tupperware.co.id	225.46K	368	7
12	1	172.17.21.205	www.detik.com	195.10K	5,965	31
12	1	172.17.21.205	www.kaskus.co.id	194.53K	6,864	11
12	1	172.17.21.205	channelbox.iklanbaris.detik.com	182.69K	13,575	73
12	1	172.17.21.205	static.ak.facebook.com	167.50K	27	6
12	1	172.17.21.205	beta.newopenx.detik.com	118.60K	669	19
12	1	172.17.21.205	1.bp.blogspot.com	112.69K	1,469	5
12	1	172.17.21.205	platform.twitter.com	103.94K	82	3
12	1	172.17.21.205	ajax.googleapis.com	100.15K	65	3

Gambar 4.5 Data Hasil Impor

Pada tabel 4.5 belum memiliki primary key sehingga perlu ditambahkan agar memudahkan untuk pengolahan data seperti yang telihat pada gambar 4.6.

#	Nama	Jenis	Penyortiran	Atribut	Kosong	Bawaan	Ekstra
1	id	bigint(20)			Tidak	*Tidak ada*	AUTO_INCREMENT
2	bulan	int(11)			Tidak	*Tidak ada*	
3	tanggal	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*	
4	ip	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*	
5	site	varchar(500)	latin1_swedish_ci		Tidak	*Tidak ada*	
6	bytes	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*	
7	millisec	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*	
8	konek	varchar(20)	latin1_swedish_ci		Tidak	*Tidak ada*	

Gambar 4.6 Struktur Tabel yang telah memiliki Primary Key

id	bulan	tanggal	ip	site	bytes	millisec	konek
17	12	1	172.17.21.205	dlcdnet.asus.com	3.48G	2,445,714	243
18	12	1	172.17.21.205	s.kaskus.id	3.59M	7,879	234
19	12	1	172.17.21.205	images.detik.com	1.84M	1,13	76
20	12	1	172.17.21.205	cdn.kaskus.com	1.74M	7,532	19
21	12	1	172.17.21.205	connect.facebook.net	632.00K	232	12
22	12	1	172.17.21.205	newopenx.detik.com	557.22K	10,027	97
23	12	1	172.17.21.205	detik.net.id	509.62K	2,204	64
24	12	1	172.17.21.205	4.bp.blogspot.com	423.54K	2,073	11
25	12	1	172.17.21.205	cdn.maskool.in	316.76K	1,958	5
26	12	1	172.17.21.205	www.google-analytics.com	278.86K	5,025	67
27	12	1	172.17.21.205	www.tupperware.co.id	225.46K	368	7
28	12	1	172.17.21.205	www.detik.com	195.10K	5,965	31
29	12	1	172.17.21.205	www.kaskus.co.id	194.53K	6,864	11
30	12	1	172.17.21.205	channelbox.iklanbaris.detik.com	182.69K	13,575	73
31	12	1	172.17.21.205	static.ak.facebook.com	167.50K	27	6
32	12	1	172.17.21.205	beta.newopenx.detik.com	118.60K	669	19
33	12	1	172.17.21.205	1.bp.blogspot.com	112.69K	1,469	5
34	12	1	172.17.21.205	platform.twitter.com	103.94K	82	3

Gambar 4.7 Tabel yang telah memiliki Primary Key

4.1.2. Data Cleaning

Data cleaning diperlukan untuk menghilangkan data yang kosong maupun data yang redundan. Dalam hal ini data cleaning menggunakan query pada MySQL. Data semula berjumlah 95.369, setelah dilakukan pembersihan, maka data yang bisa digunakan sejumlah 88.803

Tabel 4.1 Pembersihan Data Uji dengan Query

No	Keterangan Cleaning Data	Script SQL	Jumlah Data Berkurang	Jumlah Data Uji
1.	Menghapus data ip yang kosong	select count(*) FROM sites WHERE ip=''	1.730	93.639
2	Menghapus data bytes yang kosong	SELECT count(*) FROM `sites` WHERE bytes=''	24	93.615
3.	Menghapus data sites yang kosong	DELETE FROM `sites` WHERE millisec =''	19	93.596

No	Keterangan Cleaning Data	Script SQL	Jumlah Data Berkurang	Jumlah Data Uji
4.	Menghapus data bytes yang bernilai 0	SELECT count(*) FROM ` sites` WHERE bytes='0'	231	93.365
5.	Menghapus fitur data millisec yang bernilai 0	SELECT * FROM `sites` WHERE millisec='0'	4.562	88.803

4.1.3. Transformasi Data

Fitur yang masih berupa huruf harus dirubah menjadi numerik. Hal ini perlu dilakukan agar data yang dimiliki sebuah fitur dapat diproses ke tahapan selanjutnya. Adapun fitur yang harus ditransformasi menjadi numerik adalah data url website (site) dan durasi (millisec), ukuran data (bytes) dan jumlah koneksi (konek).

4.1.3.1. Transformasi data pada Fitur Ukuran Data

Ukuran data yang diakses oleh pengguna/client bervariasi ada yang berukuran bytes, kilo bytes maupun mega bytes. Untuk itu perlu distandarkan menjadi bytes. Contoh ukuran data dapat dilihat pada tabel 4.2:

Tabel 4.2 Contoh Ukuran Data

No	URL	Ukuran Data
1	www.google-analytics.com	278.86K
2	www.tupperware.co.id	225.46K
3	www.goal.com	2.13M
4	dlcdnet.asus.com	3.48G
5	www.mediafire.com	2.36M
6	www.kotakgame.com	1.70M
7	www.netmarble.co.id	1.00M

4.1.3.2 Transformasi pada fitur URL

Dalam penelitian ini URL yang diakses oleh pengguna/client ditransformasikan kedalam bentuk numerik berdasarkan domain URL dan kata yang terkandung URL tersebut. Adapun tahapan dalam proses pengkategorian URL adalah sebagai berikut:

1. Kategorisasi Website

   Setiap alamat url yang diakses oleh pengguna dikategorikan berdasarkan kategori yang telah ditetapkan sebelumnya. Adapun kategori yang telah disimpan dalam database sejumlah 8 kategori yakni:

   - Pemerintahan
   - Pendidikan/iptek/science
   - Streaming
   - Email
   - Blog/online shop
   - Media sosial
   - Berita dan
   - Pornografi.

   Jika URL yang diakses oleh pengguna tidak terdapat dalam kategori yang telah ditetapkan, maka URL tersebut dikategorikan dalam kategori lain-lain.

2. Tokenisasi alamat URL

   Dalam proses tokenisasi terdapat beberapa tahapan yang dilakukan yakni:

   a. Pemisahan setiap kata yang berada dalam URL

      Sebagai contoh jika ada user yang mengakses alamat www.facebook.com, maka alamat url akan dipisahkan (parsing) sehingga didapatkan 3 kata yaitu www, facebook dan com.

   b. Penyimpanan Kata

      Setiap kata yang telah dipisahkan disimpan dalam tabel keyword.

   c. Pemberian kategori dari setiap kata

      Administrator selaku pengelola aplikasi akan memberikan kategori dari masing-masing kata-kata tersebut bahwa facebook termasuk dalam kategori media sosial.

3. Perbandingan

   Setelah setiap kata dipisahkan, maka kata-kata tersebut dibandingkan kategorinya. 8 Kategori selain kategori lain-lain bernilai lebih tinggi dibandingkan dengan kategori lain-lain. Sehingga www.facebook.com termasuk dalam kategori media sosial dan bukan kategori lain-lain.

Tabel 4.3 Contoh Kategorisasi Website

No	Alamat URL	Kategori
1	www.googleadservices.com	Pendidikan/iptek/science
2	www.lsf.go.id	Pemerintahan
3	www.cincopa.com	Lain-lain
4	emupdate.avast.com	Pendidikan/iptek/science
5	hctsd07.blogspot.com	Blog/online shop

Alamat URL pada tabel 4.3 diubah menjadi numerik agar dapat diolah ke tahapan selanjutnya. Perubahan alamat URL dapat dilihat pada tabel 4.4.

Tabel 4.4 Transformasi Alamat URL

No	Alamat URL	Nilai Kategori
1	www.googleadservices.com	2
2	www.lsf.go.id	1
3	www.cincopa.com	9
4	emupdate.avast.com	2
5	hctsd07.blogspot.com	5

4.1.4 Identifikasi URL

Dari hasil transformasi, dapat diketahui bahwa jumlah penggguna adalah sejumlah 1.275 ip address. 1 ip address mewakili 1 pengguna. Setiap URL yang diakses oleh masing-masing ip address disimpan dalam database. Sebagai contoh dalam tabel 4.5 dapat dilihat daftar URL yang diakses oleh pengguna/user/client nomor 7.

Tabel 4.5 Daftar URL yang diakses oleh Pengguna 7

No	Kategori URL	Size	Durasi
1	Pendidikan / Iptek	3480000000	2445714
2	Media Sosial	3590000	7879
3	Berita	1840000	113
4	Media Sosial	1740000	7532
5	Media Sosial	632000	232
6	Berita	557220	10027
7	Lain-lain	509620	2204
8	Blog / Online Shop	423540	2073
9	Lain-lain	316760	1958
10	Pendidikan / Iptek	278860	5025
.	.	.	.
.	.	.	.
.	.	.	.
79	Pendidikan / Iptek	306	16

Setelah diketahui daftar URL yang diakses oleh masing-masing pengguna, langkah selanjutnya adalah mencari nilai rata-rata dari nilai durasi (millisec), ukuran data (bytes) dan jumlah koneksi yang dilakukan oleh pengguna dari masing-masing kategori website dengan menggunakan perintah sql pada database mysql:

```
select ip,id_cat,avg(bytes),avg(millisec), avg(konek) from
konvertotal group by ip, id_cat
```

Hasil dari query pengelompokan berdasarkan pengguna dapat dilihat pada tabel 4.6.

Tabel 4.6 Pengelompokan berdasarkan Pengguna

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 1	Pendidikan / Iptek	749.00	59.00	1.00
Pengguna 1	Lain-lain	5,105.50	162.17	3.00
Pengguna 2	Pemerintahan	247,206.67	928.00	1.00
Pengguna 2	Media Sosial	1,775.00	34.50	3.00
Pengguna 2	Blog / Online Shop	132,500.00	609.00	10.50
Pengguna 2	Pendidikan / Iptek	40,676.14	977.71	3.29
Pengguna 2	Lain-lain	25,634.75	5,044.75	8.50
Pengguna 3	Pendidikan / Iptek	1,230,250.00	3.25	6.75
Pengguna 3	Lain-lain	469,275.00	183.00	157.50
Pengguna 4	Media Sosial	107,245.00	141.00	4.00
Pengguna 4	Berita	107,518.33	158.83	24.17
Pengguna 4	Pendidikan / Iptek	25,604.45	114.14	8.41
Pengguna 4	Lain-lain	32,779.13	175.67	4.00
Pengguna 5	Blog / Online Shop	989,360.00	573.00	12.50
Pengguna 5	Berita	1,253,060.53	149.05	86.72
Pengguna 5	Pendidikan / Iptek	56,728.50	25.05	14.65
Pengguna 5	Streaming	6,247,812.50	104.50	65.25
Pengguna 5	Lain-lain	13,787.48	85.50	3.48
Pengguna 6	Pendidikan / Iptek	21,190.17	1,697,077.33	35.00
Pengguna 7	Pemerintahan	1,545,781.00	46.40	21.60
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Media Sosial	44,080.00	59.67	2.00

Setelah diketahui URL yang diakses oleh masing-masing pengguna/ip address maka langkah selanjutnya pengguna dikelompokkan berdasarkan masing-masing kategori. Pengelompokan pengguna berdasarkan kategori berita dapat dilihat pada tabel 4.7.

Tabel 4.7 Pengelompokan Pengguna Berdasarkan Kategori Berita

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 4	Berita	107,518.33	158.83	24.17
Pengguna 5	Berita	1,253,060.53	149.05	86.72
Pengguna 7	Berita	395,854.10	193.23	14.15
Pengguna 8	Berita	1,652,719.17	92.58	85.74
Pengguna 11	Berita	2,420,000.00	140.00	184.00
Pengguna 13	Berita	469.00	126.00	1.00
Pengguna 17	Berita	339,067.50	28.00	11.25
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Berita	8,580.00	1.00	1.00

Pengelompokan pengguna yang mengakses URL berdasarkan kategori blog/online shop dapat dilihat pada tabel 4.8.

Tabel 4.8 Pengelompokan berdasarkan kategori Blog / Online Shop

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 2	Blog / Online Shop	132,500.00	609.00	10.50
Pengguna 5	Blog / Online Shop	989,360.00	573.00	12.50
Pengguna 7	Blog / Online Shop	87,368.57	236.14	1.57
Pengguna 8	Blog / Online Shop	320,884.54	1,097.19	24.86
Pengguna 10	Blog / Online Shop	189,820.00	376.00	1.00
Pengguna 11	Blog / Online Shop	3,320.00	45.00	8.00
Pengguna 19	Blog / Online Shop	67,322.50	332.50	2.17
Pengguna 20	Blog / Online Shop	1,930.00	119.00	1.00

Pengelompokan pengguna yang mengakses URL berdasarkan kategori pemerintahan dapat dilihat pada tabel 4.9.

Tabel 4.9 Pengelompokan Berdasarkan Kategori Pemerintahan

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 2	Pemerintahan	247,206.67	928.00	1.00
Pengguna 7	Pemerintahan	1,545,781.00	46.40	21.60
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Pemerintahan	53,830.00	3,429.50	13.00

Pengelompokan pengguna yang mengakses URL berdasarkan kategori media sosial dapat dilihat pada tabel 4.10.

Tabel 4.10 Pengelompokan Berdasarkan Kategori Media Sosial

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 2	Media Sosial	1,775.00	34.50	3.00
Pengguna 4	Media Sosial	107,245.00	141.00	4.00
Pengguna 7	Media Sosial	1,598,270.00	210.73	37.64
Pengguna 8	Media Sosial	262,812.75	20.75	7.50
Pengguna 10	Media Sosial	13,160.00	163.00	8.00
Pengguna 12	Media Sosial	849.00	251.33	1.33
Pengguna 17	Media Sosial	69,440.00	172.00	2.00
Pengguna 19	Media Sosial	9,202.80	83.60	1.00
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Media Sosial	44,080.00	59.67	2.00

Pengelompokan pengguna yang mengakses URL berdasarkan kategori pendidikan / ilmu pengetahuan dapat dilihat pada tabel 4.11.

Tabel 4.11 Pengelompokan Berdasarkan Kategori Pendidikan / Iptek

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 1	Pendidikan/Iptek	749.00	59.00	1.00
Pengguna 2	Pendidikan / Iptek	40,676.14	977.71	3.29
Pengguna 3	Pendidikan / Iptek	1,230,250.00	3.25	6.75
Pengguna 4	Pendidikan / Iptek	25,604.45	114.14	8.41
Pengguna 5	Pendidikan / Iptek	56,728.50	25.05	14.65
Pengguna 6	Pendidikan / Iptek	21,190.17	1,697,077.33	35.00
Pengguna 7	Pendidikan / Iptek	782,642.00	164.73	15.87
Pengguna 8	Pendidikan / Iptek	138,624.62	127.28	12.34
Pengguna 9	Pendidikan / Iptek	16,151.00	425.00	4.67
Pengguna 10	Pendidikan / Iptek	385.00	48.00	1.00
Pengguna 11	Pendidikan / Iptek	37,530.00	321.00	63.00
Pengguna 12	Pendidikan / Iptek	1,808.13	254.88	2.75
Pengguna 13	Pendidikan / Iptek	40,268.60	158.07	1.47
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Pendidikan / Iptek	102,273.00	599.00	4.75

Pengelompokan pengguna yang mengakses URL berdasarkan kategori streaming dapat dilihat pada tabel 4.12.

Tabel 4.12 Pengelompokan Berdasarkan Kategori Streaming

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 5	Streaming	6,247,812.50	104.50	65.25
Pengguna 7	Streaming	608,464.04	93.71	5.25
Pengguna 8	Streaming	2,738,272.00	58.80	19.40
Pengguna 12	Streaming	1,770.00	2.00	5.00
Pengguna 17	Streaming	33,409,143.57	82.57	68.14
Pengguna 19	Streaming	140,860.00	1.00	23.00

Pengelompokan pengguna yang mengakses URL yang tidak termasuk dalam 8 kategori yang telah ditetapkan sebelumnya dapat dilihat pada tabel 4.13.

Tabel 4.13 Pengelompokan diluar Kategori yang telah ditetapkan

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 1	Lain-lain	5,105.50	162.17	3.00
Pengguna 2	Lain-lain	25,634.75	5,044.75	8.50
Pengguna 3	Lain-lain	469,275.00	183.00	157.50
Pengguna 4	Lain-lain	32,779.13	175.67	4.00
Pengguna 5	Lain-lain	13,787.48	85.50	3.48
Pengguna 7	Lain-lain	209,619.55	176.01	13.06
Pengguna 8	Lain-lain	105,812.14	108.68	10.82
Pengguna 10	Lain-lain	5,263.75	33.75	1.50
Pengguna 12	Lain-lain	1,387.67	543.33	1.00
Pengguna 13	Lain-lain	56,390.67	628.00	3.67
Pengguna 15	Lain-lain	1,970.00	1.00	1.00
Pengguna 17	Lain-lain	8,373,496.87	148.27	64.13
Pengguna 18	Lain-lain	16,603.60	356.40	9.80
Pengguna 19	Lain-lain	76,708.87	205.57	5.40
Pengguna 20	Lain-lain	5,196.67	151.67	2.67
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Lain-lain	113,882.77	125.73	28.16

4.1.5 Normalisasi

Setelah dilakukan pengelompokan, maka setiap kategori data perlu dilakukan normalisasi menggunakan metode min-max sehingga menghasilkan data yang bisa digunakan untuk proses clustering. Data yang dinormalisasi untuk kategori berita dapat dilihat pada tabel 4.14.

Tabel 4.14 Normalisasi untuk Kategori Berita

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 4	Berita	5.38	82.29	13.53
Pengguna 5	Berita	52.25	77.25	47.37
Pengguna 7	Berita	17.18	100.00	8.12
Pengguna 8	Berita	68.61	48.17	46.84
Pengguna 11	Berita	100.00	72.59	100.00
Pengguna 13	Berita	1.00	65.38	1.00
Pengguna 17	Berita	14.85	14.91	6.55
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Berita	1.33	1.00	1.00

Data yang dinormalisasi untuk kategori blog/online shop dapat dilihat pada tabel 4.15.

Tabel 4.15 Normalisasi untuk Kategori Blog/Online Shop

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 2	Blog / Online Shop	14.09	54.07	40.41
Pengguna 5	Blog / Online Shop	100.00	50.68	48.71
Pengguna 7	Blog / Online Shop	9.57	18.98	3.37
Pengguna 8	Blog / Online Shop	32.98	100.00	100.00
Pengguna 10	Blog / Online Shop	19.84	32.14	1.00
Pengguna 11	Blog / Online Shop	1.14	1.00	30.04
Pengguna 19	Blog / Online Shop	7.56	28.05	5.84
Pengguna 20	Blog / Online Shop	14.09	54.07	40.41

Data yang dinormalisasi untuk kategori Pemerintahan dapat dilihat pada tabel 4.16.

Tabel 4.16 Normalisasi untuk Kategori Pemerintahan

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 2	Pemerintahan	13.83	26.80	1.00
Pengguna 7	Pemerintahan	100.00	1.00	100.00
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Pemerintahan	1.00	100.00	58.67

Data yang dinormalisasi untuk kategori media sosial dapat dilihat pada tabel 4.17.

Tabel 4.17 Normalisasi untuk Kategori Media Sosial

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 2	Media Sosial	1.06	6.90	6.40
Pengguna 4	Media Sosial	7.59	52.63	9.11
Pengguna 7	Media Sosial	100.00	82.57	100.00
Pengguna 8	Media Sosial	17.24	1.00	18.56
Pengguna 10	Media Sosial	1.76	62.07	19.92
Pengguna 12	Media Sosial	1.00	100.00	1.90
Pengguna 17	Media Sosial	5.25	65.94	3.70
Pengguna 19	Media Sosial	1.52	27.98	1.00
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Media Sosial	3.68	17.71	3.70

Data yang dinormalisasi untuk kategori streaming dapat dilihat pada tabel 4.18.

Tabel 4.18 Normalisasi untuk Kategori Streaming

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 5	Streaming	19.51	100.00	95.46
Pengguna 7	Streaming	2.80	89.68	1.39
Pengguna 8	Streaming	9.11	56.29	23.58
Pengguna 12	Streaming	1.00	1.96	1.00
Pengguna 17	Streaming	100.00	79.02	100.00
Pengguna 19	Streaming	1.41	1.00	29.22

Data yang dinormalisasi untuk kategori Pendidikan/Iptek dapat dilihat pada tabel 4.19.

Tabel 4.19 Normalisasi untuk Kategori Pendidikan/Iptek

Pengguna/user	Kategori	Size	Durasi	Konek
Pengguna 1	Pendidikan / Iptek	1.03	1.00	1.00
Pengguna 2	Pendidikan / Iptek	4.24	1.06	4.65
Pengguna 3	Pendidikan / Iptek	100.00	1.00	10.18
Pengguna 4	Pendidikan / Iptek	3.03	1.01	12.83
Pengguna 5	Pendidikan / Iptek	5.54	1.00	22.80
Pengguna 6	Pendidikan / Iptek	2.67	100.00	55.29
Pengguna 7	Pendidikan / Iptek	63.97	1.01	24.74
Pengguna 8	Pendidikan / Iptek	12.13	1.01	19.12
Pengguna 9	Pendidikan / Iptek	2.27	1.02	6.85

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 10	Pendidikan / Iptek	1.00	1.00	1.00
Pengguna 11	Pendidikan / Iptek	3.99	1.02	100.00
Pengguna 12	Pendidikan / Iptek	1.11	1.01	3.79
Pengguna 13	Pendidikan / Iptek	4.21	1.01	1.75
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Pendidikan / Iptek	9.20	1.03	6.99

Data yang dinormalisasi untuk kategori streaming dapat dilihat pada tabel 4.20.

Tabel 4.20 Normalisasi untuk Data diluar Kategori yang ditetapkan

Pengguna/User	Kategori	Size	Durasi	Konek
Pengguna 1	Lain-lain	1.04	4.16	2.27
Pengguna 2	Lain-lain	1.29	100.00	5.74
Pengguna 3	Lain-lain	6.53	4.57	100.00
Pengguna 4	Lain-lain	1.37	4.43	2.90
Pengguna 5	Lain-lain	1.15	2.66	2.57
Pengguna 7	Lain-lain	3.46	4.44	8.63
Pengguna 8	Lain-lain	2.23	3.11	7.21
Pengguna 10	Lain-lain	1.05	1.64	1.32
Pengguna 12	Lain-lain	1.00	11.65	1.00
Pengguna 13	Lain-lain	1.65	13.31	2.69
Pengguna 15	Lain-lain	1.01	1.00	1.00
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
Pengguna 1275	Lain-lain	2.33	3.45	18.18

4.2 Uji Coba

Uji coba dilakukan dengan membandingkan nilai Silhouette Index yang diperoleh metode Genetic K-Means Algorithm dan K-Means untuk masing-masing kategori dan ujicoba dengan melakukan perubahan probabilitas mutasi untuk masing-masing cluster disetiap kategori.

4.2.1. Perbandingan Nilai

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-Means untuk kategori Pemerintahan dapat dilihat pada tabel 4.21.

Tabel 4.21 Perbandingan Nilai Silhouette Index untuk Kategori Pemerintahan

Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.909	0.909
Cluster 2 (jarang)	0.438	0.438
Cluster 3 (sering)	0.679	0.679
Cluster 4 (sangat sering)	0.588	0.588
SI Global	**0.653**	**0.653**

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-means untuk kategori Email dapat dilihat pada tabel 4.22.

Tabel 4.22 Perbandingan Nilai Silhouette Index untuk Kategori Email

Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.852	0.379
Cluster 2 (jarang)	1.000	1.000
Cluster 3 (sering)	0.898	0.986
Cluster 4 (sangat sering)	0.982	0.984
SI Global	**0.933**	**0.837**

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-means untuk kategori Media Sosial dapat dilihat pada tabel 4.23.

Tabel 4.23 Perbandingan Nilai Silhouette Index untuk Kategori Media Sosial

Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.593	0.593
Cluster 2 (jarang)	0.934	0.934
Cluster 3 (sering)	0.720	0.720
Cluster 4 (sangat sering)	0.922	0.922
SI Global	0.792	0.792

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-means untuk kategori Blog / Online Shop dapat dilihat pada tabel 4.24.

Tabel 4.24 Perbandingan Nilai Silhouette Index untuk Kategori Blog/Online Shop

Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.853	0.967
Cluster 2 (jarang)	0.420	0.428
Cluster 3 (sering)	0.968	0.866
Cluster 4 (sangat sering)	0.869	0.675
SI Global	**0.778**	**0.734**

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-means untuk kategori Berita dapat dilihat pada tabel 4.25.

Tabel 4.25 Perbandingan Nilai Silhouette Index untuk Kategori Berita

Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.570	0.563
Cluster 2 (jarang)	0.555	0.897
Cluster 3 (sering)	0.893	0.370
Cluster 4 (sangat sering)	0.548	0.596
SI Global	**0.641**	**0.607**

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-means untuk kategori Pendidikan/Iptek dapat dilihat pada tabel 4.26.

Tabel 4.26 Perbandingan Nilai Silhouette Index untuk Kategori Pendidikan/Iptek

Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.559	0.559
Cluster 2 (jarang)	0.440	0.440
Cluster 3 (sering)	0.932	0.932
Cluster 4 (sangat sering)	0.079	0.079
SI Global	0.503	0.503

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-means untuk kategori streaming dapat dilihat pada tabel 4.27.

Tabel 4. 27 Perbandingan Nilai Silhouette Index untuk Kategori Streaming

Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.919	0.919
Cluster 2 (jarang)	-0.044	-0.044
Cluster 3 (sering)	0.506	0.506
Cluster 4 (sangat sering)	0.678	0.678
SI Global	**0.515**	**0.515**

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-means data diluar kategori yang telah ditetapkan dapat dilihat pada tabel 4.28.

Tabel 4.28 Perbandingan Nilai Silhouette Index diluar Kategori yang telah ditetapkan

Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.507	0.576
Cluster 2 (jarang)	0.680	0.680
Cluster 3 (sering)	0.562	0.470
Cluster 4 (sangat sering)	0.894	0.902
SI Global	**0.661**	**0.657**

Nilai Silhouette Index yang dihasilkan oleh metode Genetic K-Means Algorithm dan K-means untuk kategori pornografi dapat dilihat pada tabel 4.29.

Tabel 4.29 Perbandingan Nilai Silhouette Index untuk Kategori Pornografi

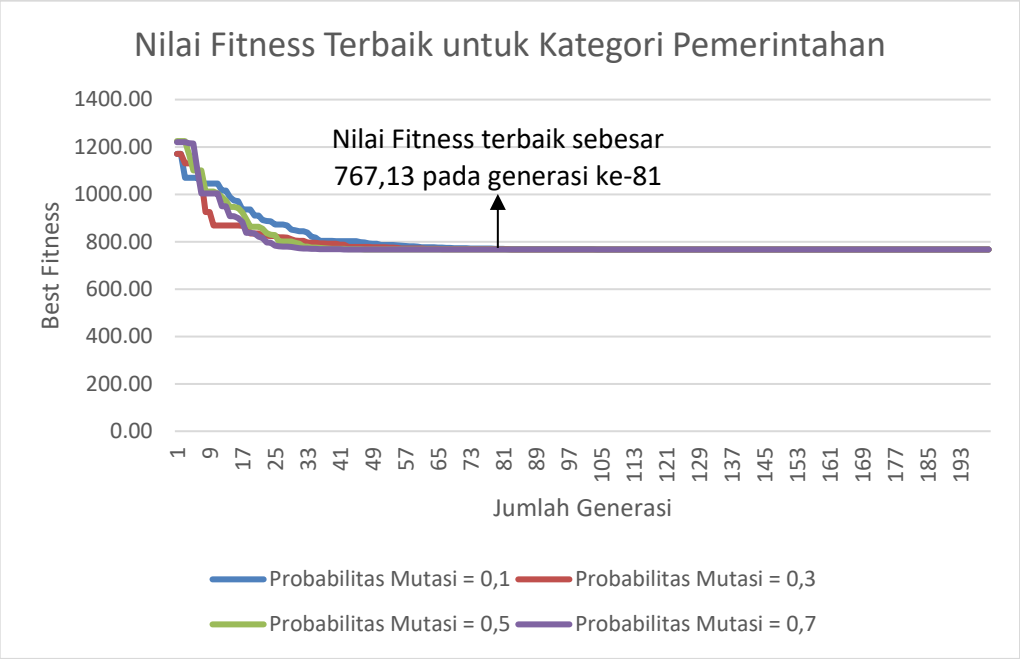
Cluster	Genetic K-Means Algorithm	K-means
Cluster 1(sangat jarang)	0.443	0.443
Cluster 2 (jarang)	0.981	0.981
Cluster 3 (sering)	1.000	1.000
Cluster 4 (sangat sering)	0.912	0.912
SI Global	**0.834**	**0.834**

Nilai Silhouette Index dari Genetic K-Means Algorithm lebih baik dibandingkan dengan nilai K-Means yang berarti Genetic K-Means Algorithm lebih baik dibandingkan K-Means dalam pengelompokan data seperti nilai yang terdapat pada tabel 4.22, 4.24, 4.25 dan 4.28.

4.2.2. Perubahan Probabilitas Mutasi

Pada penelitian ini, ujicoba dilakukan dengan mengubah nilai probabilitas mutasi yang ditetapkan di awal proses Genetic K-Means Algorithm. Pada ujicoba kali ini, jumlah populasi = 1000, jumlah generasi/iterasi = 200 dan probabilitas cross over = 0,9.

Nilai Fitness terbaik untuk kategori pemerintahan sebesar 767.13 pada generasi ke-81 untuk perubahan probabilitas mutasi sebesar 0.1, 0.3, 0.5, 0.7 seperti yang terlihat pada gambar 4.8.

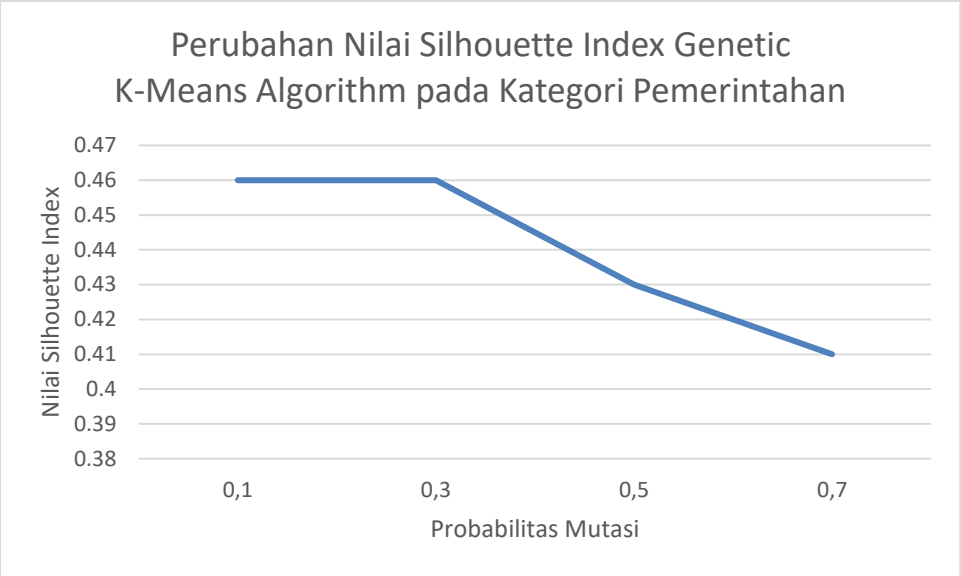


Gambar 4.8 Best Fitness pada Kategori Pemerintahan

Perubahan nilai Silhouette Index yang diakibatkan oleh perubahan probabilitas mutasi pada kategori pemerintahan dapat dilihat pada tabel 4.30 dan gambar 4.9.

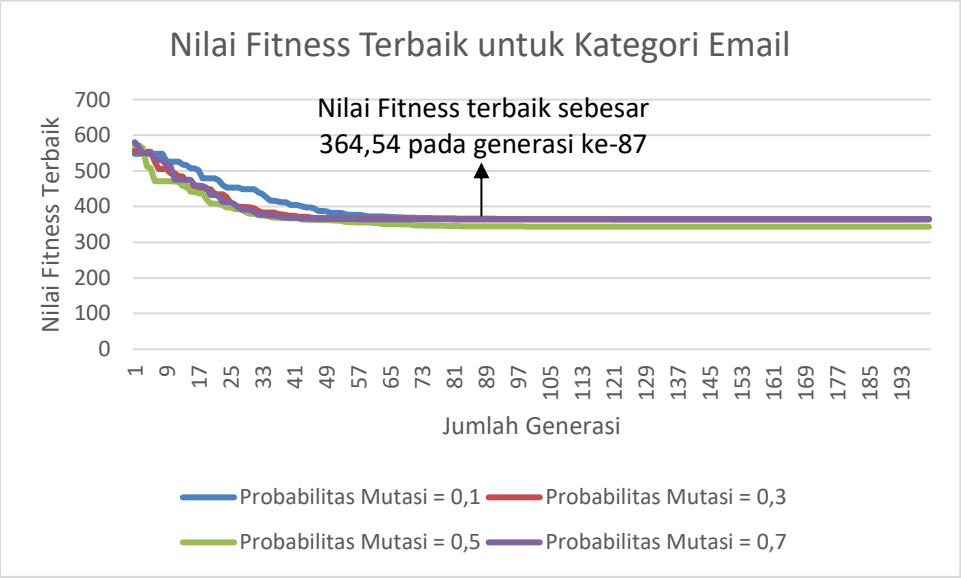
Tabel 4.30 Nilai Silhouette Index pada Kategori Pemerintahan

Probabilitas Mutasi	Nilai Silhouette Index Genetic K-Means Algorithm
0,1	0.46
0,3	0.46
0,5	0.43
0,7	0.41



Gambar 4.9 Nilai Silhouette Index pada Kategori Pemerintahan

Nilai Fitness terbaik untuk kategori email sebesar 364,64 pada generasi ke-87 untuk perubahan probabilitas mutasi sebesar 0.1, 0.3, 0.5, 0.7 seperti yang terlihat pada gambar 4.10.

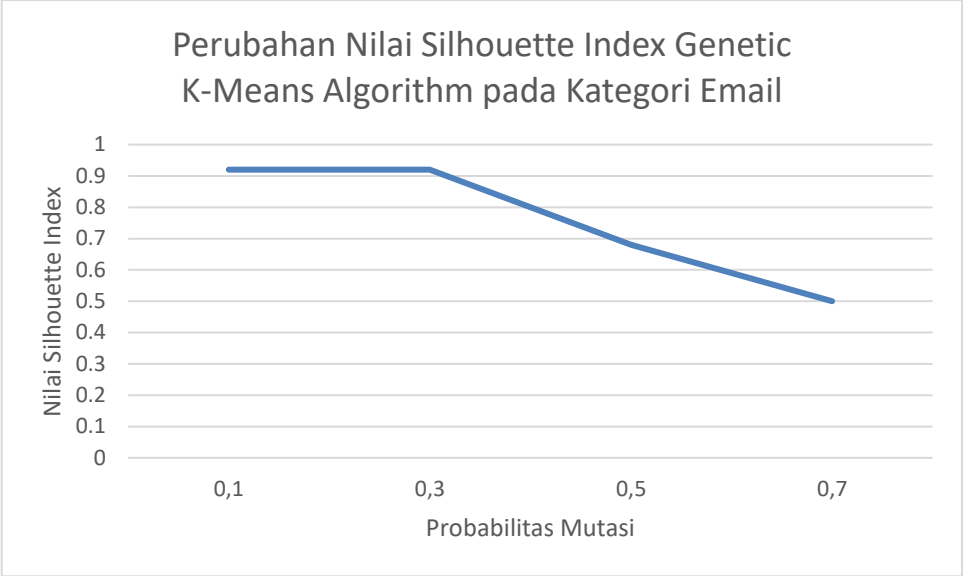


Gambar 4.10 Best Fitness pada Kategori Email

Perubahan nilai Silhouette Index yang diakibatkan oleh perubahan probabilitas mutasi pada kategori email dapat dilihat pada tabel 4.31 dan gambar 4.11.

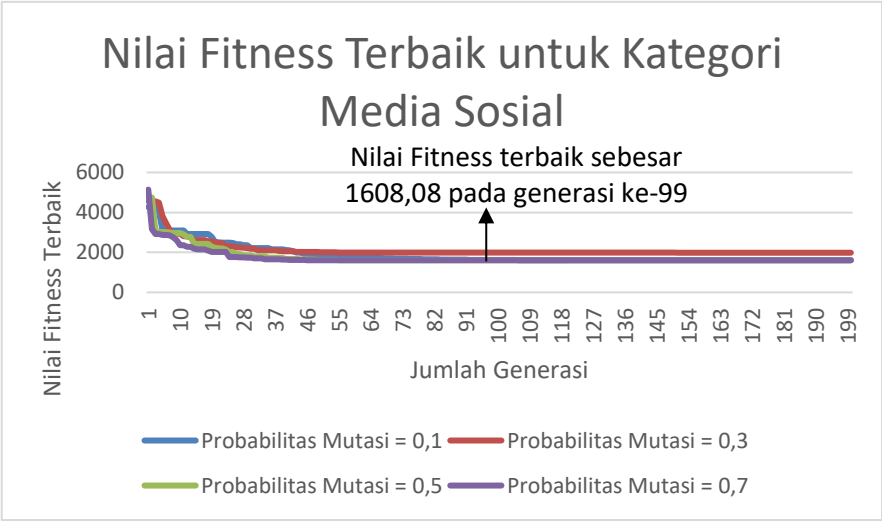
Tabel 4.31Nilai Silhouette Index pada Kategori Email

Probabilitas Mutasi	Nilai Silhouette Index Genetic K-Means Algorithm
0,1	0.92
0,3	0.92
0,5	0.68
0,7	0.50



Gambar 4.11 Nilai Silhouette Index pada Kategori Email

Nilai Fitness terbaik untuk kategori media sosial sebesar 1608,08 pada generasi ke-99 untuk perubahan probabilitas mutasi sebesar 0.1, 0.3, 0.5, 0.7 seperti yang terlihat pada gambar 4.12.

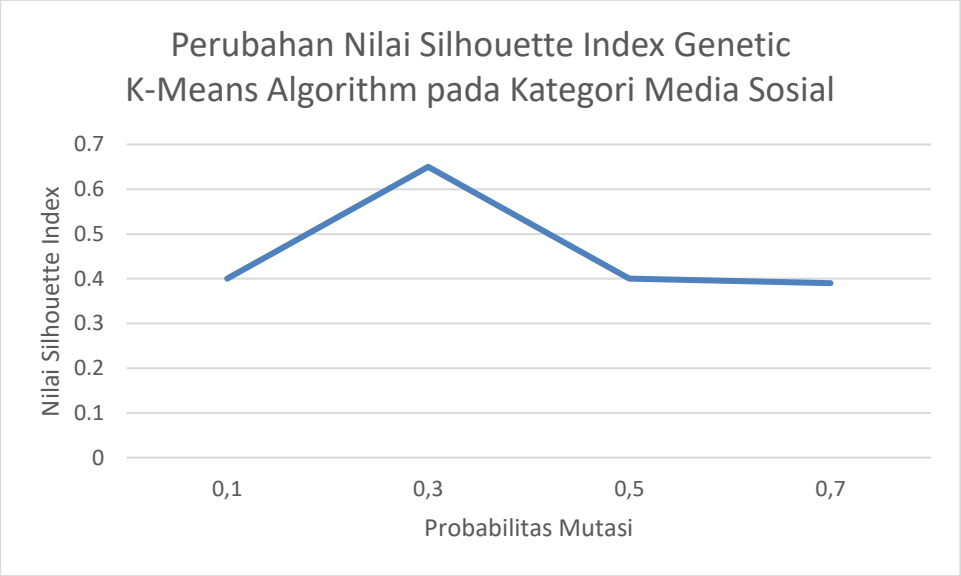


Gambar 4.12 Best Fitness pada Kategori Email

Perubahan nilai Silhouette Index yang diakibatkan oleh perubahan probabilitas mutasi pada kategori media sosial dapat dilihat pada tabel 4.32 dan gambar 4.13.

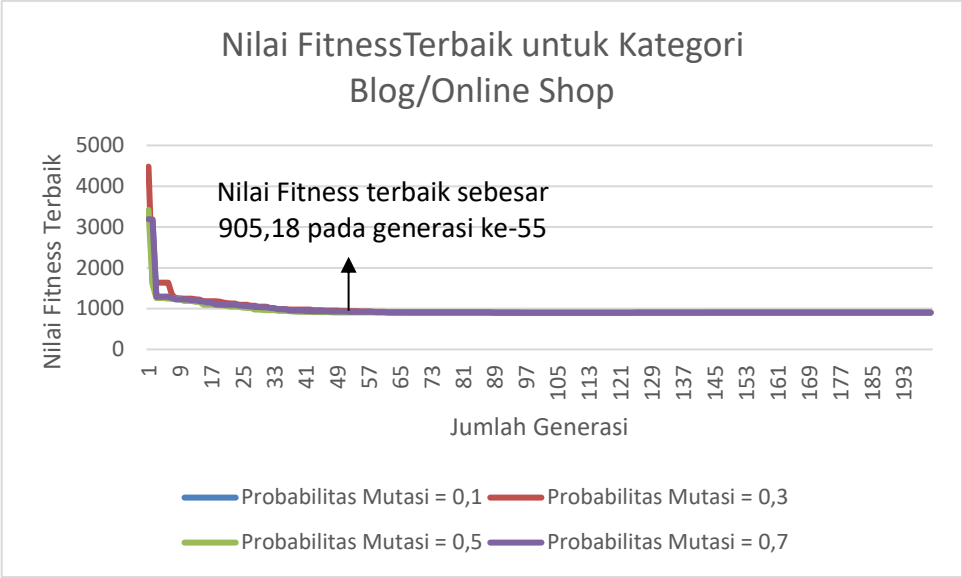
Tabel 4.32 Nilai Silhouette Index pada Kategori Media Sosial

Probabilitas Mutasi	Nilai Silhouette Index Genetic K-Means Algorithm
0,1	0.40
0,3	0.65
0,5	0.40
0,7	0.39



Gambar 4.13 Nilai Silhouette Index pada Kategori Media Sosial

Nilai Fitness terbaik untuk kategori blog/online shop sebesar 905,18 pada generasi ke-55 untuk perubahan probabilitas mutasi sebesar 0.1, 0.3, 0.5, 0.7 seperti yang terlihat pada gambar 4.14.

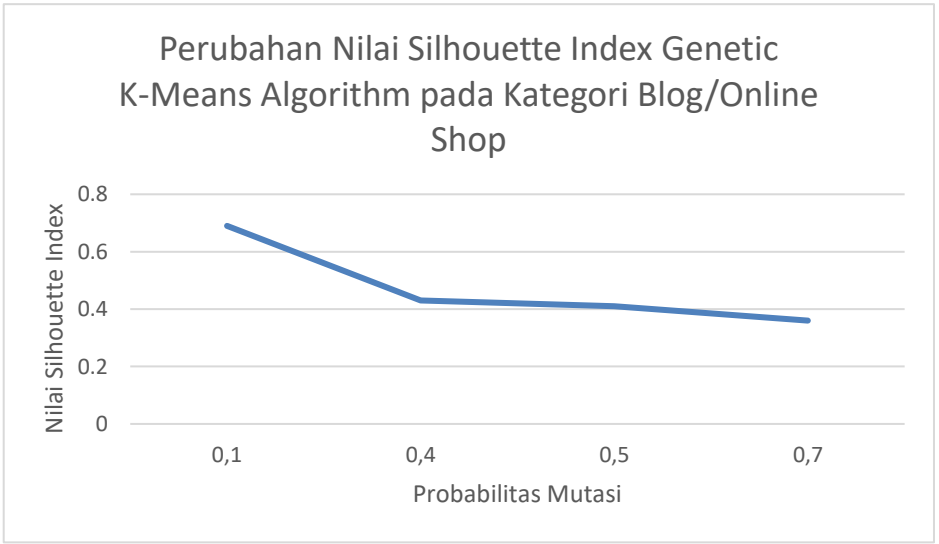


Gambar 4.14 Best Fitness pada Kategori Blog/Online Shop

Perubahan nilai Silhouette Index yang diakibatkan oleh perubahan probabilitas mutasi pada kategori blog/online shop dapat dilihat pada tabel 4.33 dan gambar 4.15.

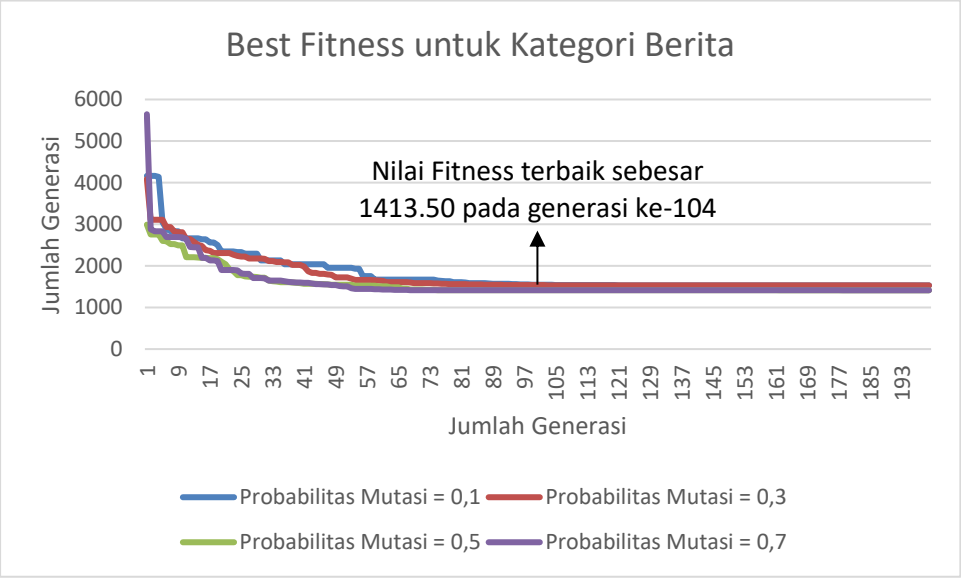
Tabel 4.33 Nilai Silhouette Index pada Kategori Blog/Online Shop

Probabilitas Mutasi	**Nilai Silhouette Index Genetic K-Means Algorithm**
0,1	0.69
0,4	0.43
0,5	0.41
0,7	0.36



Gambar 4.15 Nilai Silhouette Index pada Kategori Blog/Online Shop

Nilai Fitness terbaik untuk kategori berita sebesar 1413.50 pada generasi ke-104 untuk perubahan probabilitas mutasi sebesar 0.1, 0.3, 0.5, 0.7 seperti yang terlihat pada gambar 4.16.

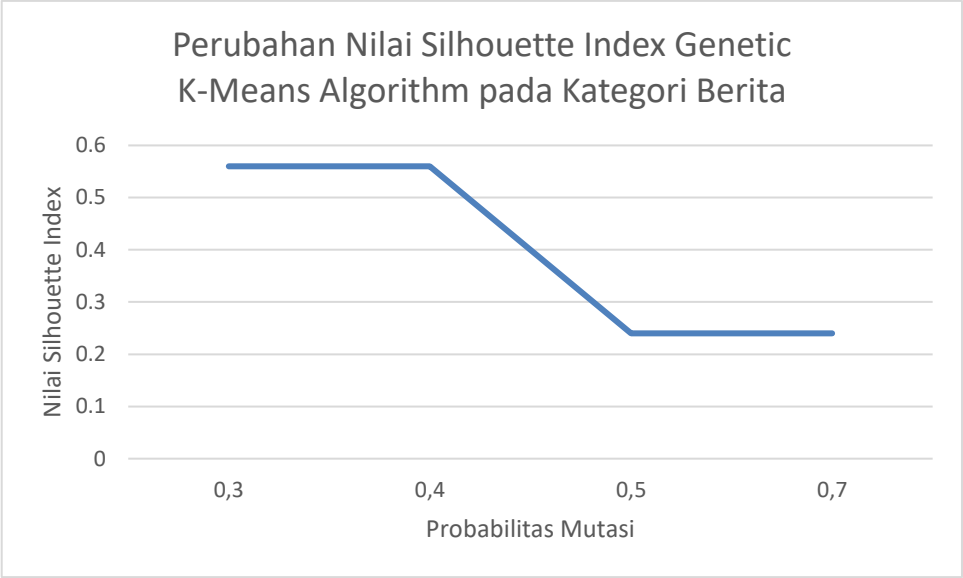


Gambar 4.16 Best Fitness pada Kategori Blog/Online Shop

Perubahan nilai Silhouette Index yang diakibatkan oleh perubahan probabilitas mutasi pada kategori berita dapat dilihat pada tabel 4.34 dan gambar 4.17.

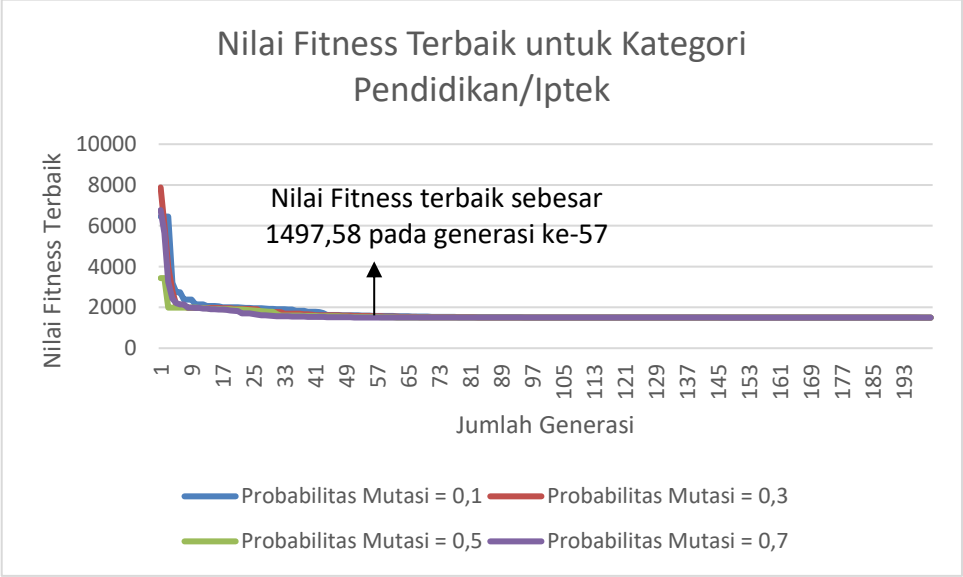
Tabel 4.34 Nilai Silhouette Index pada Kategori Berita

Probabilitas Mutasi	**Nilai Silhouette Index Genetic K-Means Algorithm**
0,3	0.56
0,4	0.56
0,5	0.24
0,7	0.24



Gambar 4.17 Nilai Silhouette Index pada Kategori Berita

Nilai Fitness terbaik untuk kategori pendidikan/iptek sebesar 1497,58 pada generasi ke-57 untuk perubahan probabilitas mutasi sebesar 0.1, 0.3, 0.5, 0.7 seperti yang terlihat pada gambar 4.18.

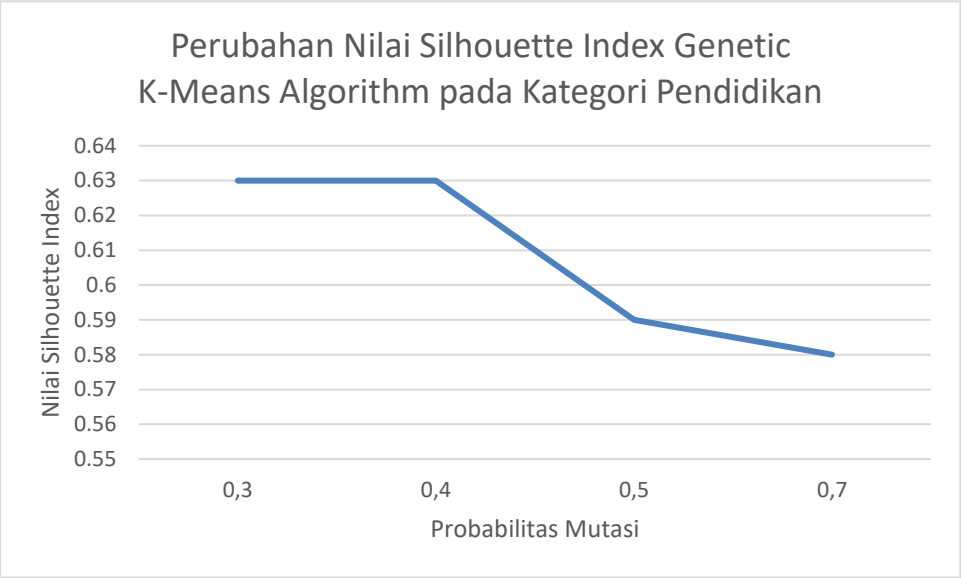


Gambar 4.18 Best Fitness pada Kategori Pendidikan/Iptek

Perubahan nilai Silhouette Index yang diakibatkan oleh perubahan probabilitas mutasi pada kategori pendidikan/iptek dapat dilihat pada tabel 4.35 dan gambar 4.19.

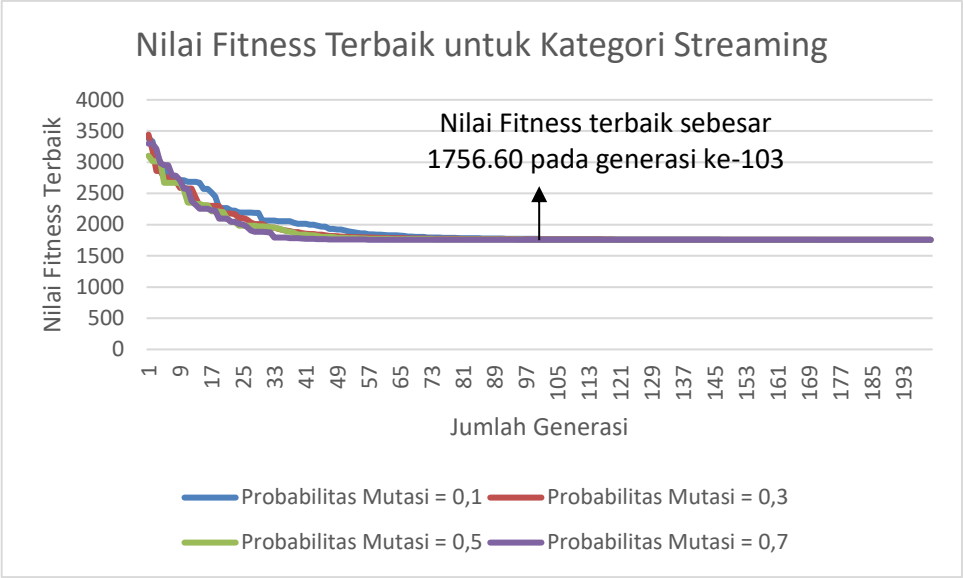
Tabel 4.35 Nilai Silhouette Index pada Kategori Pendidikan/Iptek

Probabilitas Mutasi	Nilai Silhouette Index Genetic K-Means Algorithm
0,3	0.63
0,4	0.63
0,5	0.59
0,7	0.58



Gambar 4.19 Nilai Silhouette Index pada Kategori Pendidikan/Iptek

Nilai Fitness terbaik untuk kategori streaming sebesar 1756,60 pada generasi ke-103 untuk perubahan probabilitas mutasi sebesar 0.1, 0.3, 0.5, 0.7 seperti yang terlihat pada gambar 4.20.

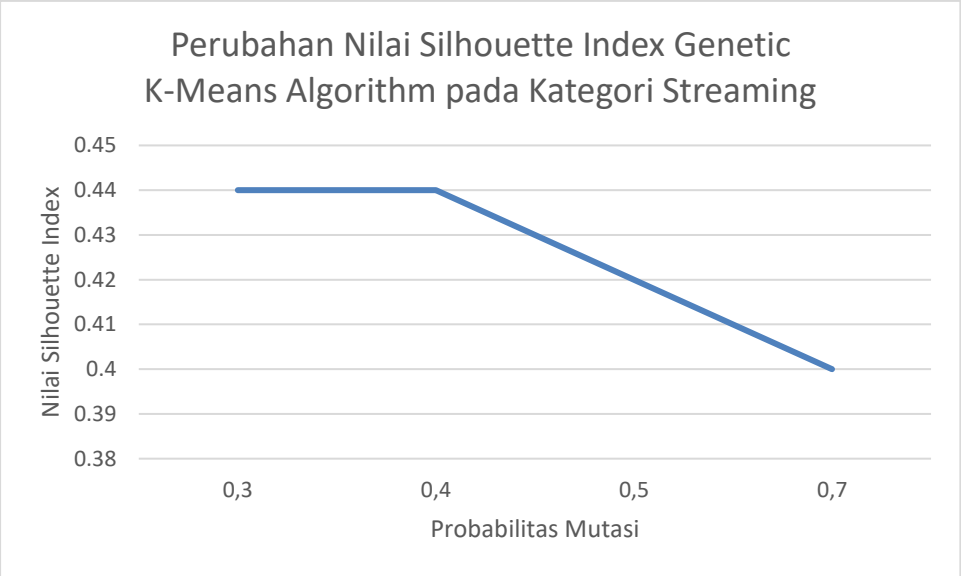


Gambar 4.20 Best Fitness pada Kategori Streaming

Perubahan nilai Silhouette Index yang diakibatkan oleh perubahan probabilitas mutasi pada kategori streaming dapat dilihat pada tabel 4.36 dan gambar 4.21.

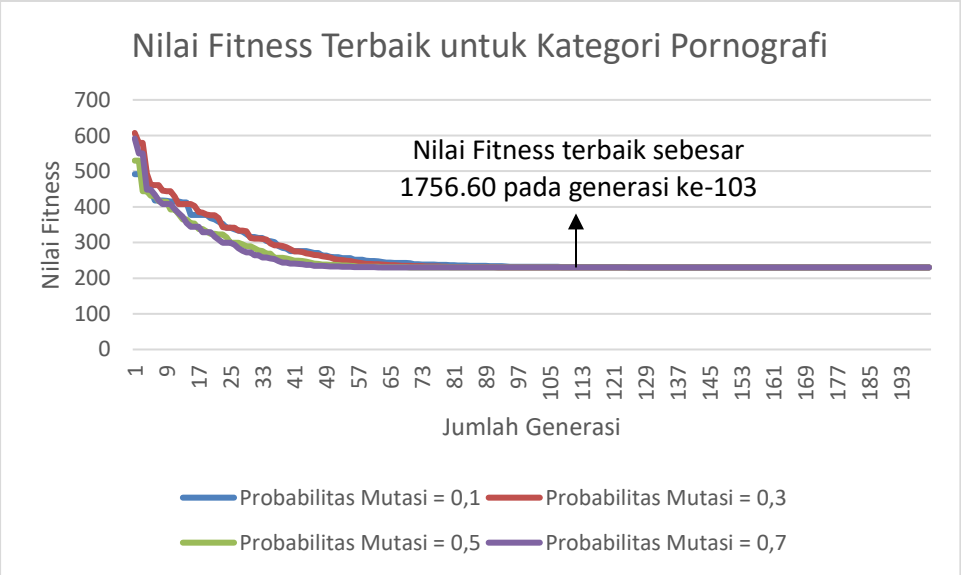
Tabel 4.36 Nilai Silhouette Index pada Kategori Streaming

Probabilitas Mutasi	**Nilai Silhouette Index Genetic K-Means Algorithm**
0,3	0.44
0,4	0.44
0,5	0.42
0,7	0.40



Gambar 4.21 Nilai Silhouette Index pada Kategori Streaming

Nilai Fitness terbaik untuk kategori pornografi sebesar 1756.60 pada generasi ke-103 untuk perubahan probabilitas mutasi sebesar 0.1, 0.3, 0.5, 0.7 seperti yang terlihat pada gambar 4.22.

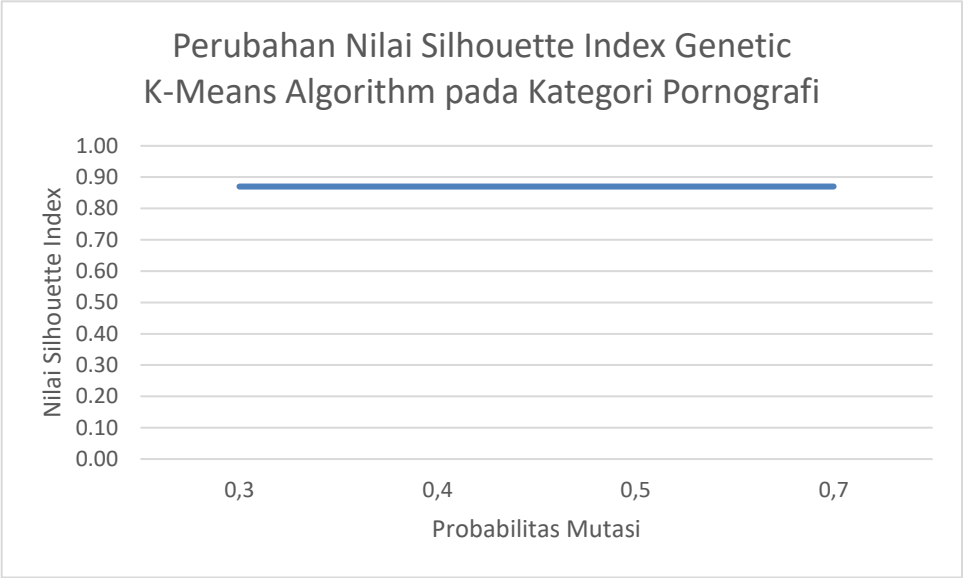


Gambar 4.22 Best Fitness pada Kategori Pornografi

Perubahan nilai Silhouette Index yang diakibatkan oleh perubahan probabilitas mutasi pada kategori pornografi dapat dilihat pada tabel 4.37 dan gambar 4.23.

Tabel 4.37 Nilai Silhouette Index pada Kategori Pornografi

Probabilitas Mutasi	Nilai Silhouette Index Genetic K-Means Algorithm
0,3	0.87
0,4	0.87
0,5	0.87
0,7	0.87



Gambar 4.23 Nilai Silhouette Index pada Kategori Pornografi

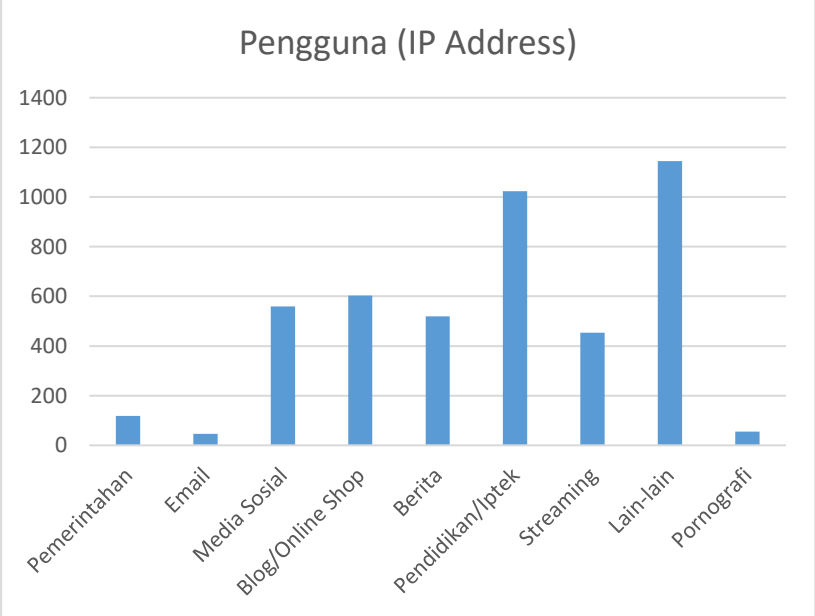
Dari tabel 4.21 sampai tabel 4.28 dapat disimpulkan bahwa semakin tinggi nilai probabilitas mutasi, maka nilai Silhouette Index yang didapatkan semakin kecil yang mencerminkan buruknya hasil cluster seiring dengan pertambahan nilai probabilitas mutasi.

4.2.3. Jumlah Pengguna

Jumlah seluruh pengguna (ip address) adalah 1.275 dan jumlah pengguna berdasarkan kategori dapat dilihat pada tabel 4.38.

Tabel 4.38 Jumlah pengguna

No	Kategori	Jumlah Pengguna (Ip Address)
1	Pemerintahan	118
2	Email	46
3	Media Sosial	560
4	Blog/Online shop	603
5	Berita	519
6	Pendidikan/Iptek	1023
7	Streaming	453
8	Lain-lain	1145
9	Pornografi	55



Gambar 4.24 Jumlah Pengguna

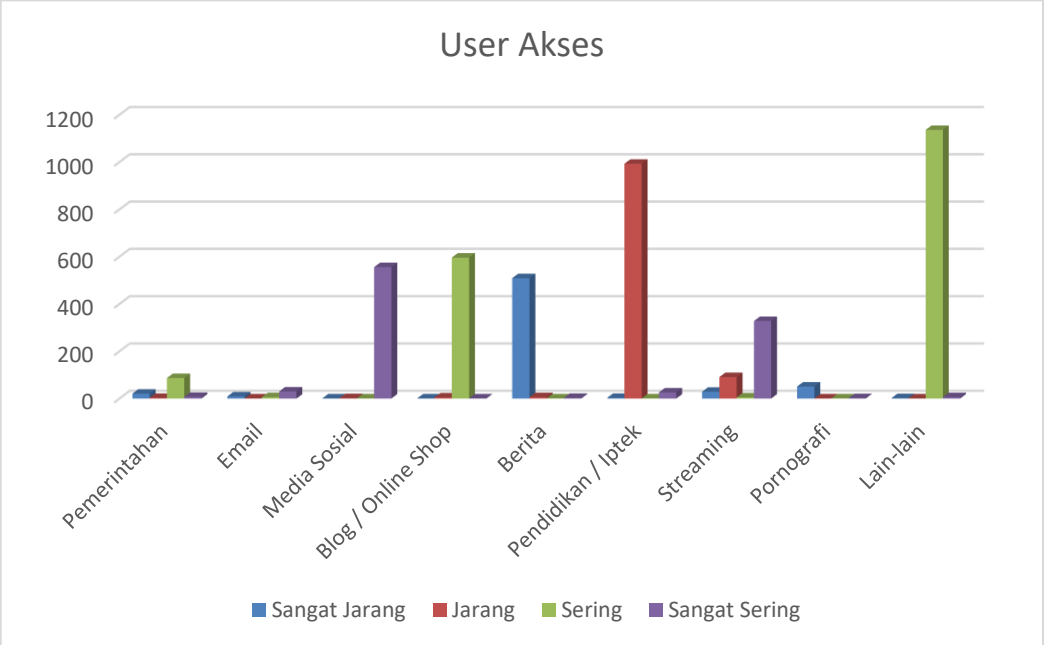
Dari tabel 4.29 dapat diketahui jumlah pengguna. Setiap data pada masing-masing kategori dicluster menggunakan metode Genetic K-Means Algorithm dan K-Means untuk mengetahui tingkat minat/kecenderungan pengguna terhadap masing-masing kategori yang telah ditetapkan.

4.2.4. Pengkategorian Akses User

Setelah diketahui jumlah masing-masing pengguna untuk masing-masing kategori sepeerti yang terlihat pada tabel 4.38 maka diperlukan rincian pengguna masing-masing kategori seperti yang dapat dilihat pada tabel 4.39.

Tabel 4.39 Pengkategorian Akses User

Kategori	Sangat Jarang (1)	Jarang (2)	Sering (3)	Sangat Sering (4)
Pemerintahan	21	3	87	7
Email	9	1	6	30
Media Sosial	1	2	1	556
Blog / Online Shop	1	5	596	1
Berita	509	6	1	3
Pendidikan / Iptek	3	993	1	26
Streaming	29	91	5	328
Pornografi	51	1	1	2
Lain-lain	2	1	1136	6



Gambar 4.25 User Akses

Contoh perilaku pengguna (ip adress) dapat dilihat pada matriks korespondensi yang dapat dilihat pada tabel 4.40.

Dari tabel 4.39 maupun gambar 4.25 dapat diketahui bahwa 3 kategori website yang diminati oleh pengguna web adalah sebagai berikut:

1. Pendidikan / ilmu pengetahuan dan teknologi
2. Blog / online shop
3. Media sosial

Untuk mengetahui minat akses masing-masing pengguna untuk mengakses website berdasarkan masing-masing kategori dapat dilihat pada tabel 4.40. Dalam matrik korespondensi terdapat nilai-nilai yang menunjukkan tingkat minat dari masing-masing pengguna:

1. Nilai 1 : sangat jarang
2. Nilai 2 : jarang
3. Nilai 3 : sering
4. Nilai 4 : sangat sering / sangat berminat

Tabel 4.40 Matrik Korespondensi

Pengguna	Pemerintahan	Email	Media Sosial	Blog/online Shop	Berita	Pendidikan / Iptek	Streaming	Lain-lain	Pornografi
Pengguna 1			4		1	2	2	3	
Pengguna 2			4			2	4	3	
Pengguna 3			4	3		2	4	3	
Pengguna 4				3		2		3	
Pengguna 5	3		4	3	1	2		3	
Pengguna 6			4	3	1	2		3	
Pengguna 7			4	4	4	2	2	3	
Pengguna 8			4	3	1	2	4	4	
Pengguna 9	3	1	4	3	1	2	4	3	
Pengguna 10	3		4	3	1	2		3	1

Dari tabel 4.40 dapat diketahui bahwa:

1. Pengguna 1 :
    - Sangat sering mengakses situs dengan kategori media sosial
    - Sangat jarang mengakses situs dengan kategori berita
    - Jarang mengakses situs dengan kategori pendidikan/iptek

- Jarang mengakses situs dengan kategori streaming

2. Pengguna 2 :
    - Sangat sering mengakses situs dengan kategori media sosial
    - Jarang mengakses situs dengan kategori pendidikan/iptek
    - Sangat sering mengakses situs dengan kategori streaming
3. Pengguna 3 :
    - Sangat sering mengakses situs dengan kategori media sosial
    - Sering mengakses situs dengan kategori blog / online shop
    - Jarang mengakses situs dengan kategori pendidikan/iptek
    - Sangat sering mengakses situs dengan kategori streaming
4. Pengguna 4 :
    - Sering mengakses situs dengan kategori blog / online shop
    - Jarang mengakses situs dengan kategori pendidikan/iptek
5. Pengguna 5 :
    - Sering mengakses situs dengan kategori pemerintahan
    - Sangat sering mengakses situs dengan kategori media sosial
    - Sering mengakses situs dengan kategori blog / online shop
    - Sangat jarang mengakses situs dengan kategori berita
    - Jarang mengakses situs dengan kategori pendidikan/iptek
6. Pengguna 6 :
    - Sangat sering mengakses situs dengan kategori media sosial
    - Sering mengakses situs dengan kategori blog / online shop
    - Sangat jarang mengakses situs dengan kategori berita
    - Jarang mengakses situs dengan kategori pendidikan/iptek
7. Pengguna 7 :
    - Sangat sering mengakses situs dengan kategori media sosial
    - Sangat sering mengakses situs dengan kategori blog / online shop
    - Sangat sering mengakses situs dengan kategori berita
    - Jarang mengakses situs dengan kategori pendidikan/iptek
8. Pengguna 8 :
    - Sangat sering mengakses situs dengan kategori media sosial

- Sering mengakses situs dengan kategori blog / online shop
- Sangat jarang mengakses situs dengan kategori berita
- Jarang mengakses situs dengan kategori pendidikan/iptek
- Sangat sering mengakses situs dengan kategori streaming

9. Pengguna 9 :
    - Sering mengakses situs dengan kategori pemerintahan
    - Sangat jarang mengakses situs dengan kategori email
    - Sangat sering mengakses situs dengan kategori media sosial
    - Sering mengakses situs dengan kategori blog / online shop
    - Sangat jarang mengakses situs dengan kategori berita
    - Jarang mengakses situs dengan kategori pendidikan/iptek
    - Sangat sering mengakses situs dengan kategori streaming
10. Pengguna 10 :
    - Sering mengakses situs dengan kategori pemerintahan
    - Sangat sering mengakses situs dengan kategori media sosial
    - Sering mengakses situs dengan kategori blog / online shop
    - Sangat jarang mengakses situs dengan kategori berita
    - Jarang mengakses situs dengan kategori pendidikan/iptek
    - Sangat jarang mengakses situs dengan kategori pornografi.

BAB 5
KESIMPULAN DAN SARAN

5.1 Kesimpulan

Kesimpulan yang dapat diambil dari penelitian ini adalah sebagai berikut.

1. Nilai Silhouette Index yang didapatkan oleh Genetic K-Means Algorithm menunjukkan bahwa terdapat perbaikan kualitas pembentukan cluster sebesar 28,71% lebih baik dibandingkan dengan K-means. Hal ini berarti Genetic K-Means Algorithm bisa mendapatkan cluster yang lebih homogen dan memiliki heterogenitas yang antar clusternya dibandingkan dengan K-means.
2. Hasil pengujian menunjukkan bahwa terdapat 2 kategori dari 8 kategori yang diminati oleh pengguna website yakni kategori blog/online shop dengan 596 pengguna yang berminat/sering mengunjungi website berbasis blog/online shop dan kategori media sosial dengan 556 pengguna yang sangat berminat/sangat sering mengunjungi website berbasis media sosial.

5.2 Saran

Adapun saran yang bisa diberikan berdasarkan hasil yang didapat dari penelitian ini adalah sebagai berikut.

1. Kategori lain-lain mempunyai anggota paling banyak karena keyword terlalu general. Untuk tahap penelitian lebih lanjut, membutuhkan tambahan keyword yang lebih spesifik untuk mendapatkan kategori yang lebih spesifik juga.
2. Perlunya dilakukan pembatasan akses terhadap website-website yang berbasis media sosial dan blog / online shop.

Halaman ini sengaja dikosongkan

DAFTAR PUSTAKA

[1] Agusta, Y., 2007. K-Means-Penerapan, Permasalahan dan Metode Terkait. Jurnal Sistem dan Informatika, 3(1), pp.47-60.

[2] Barakbah, A.R., Fariza, A. and Setiowati, Y., 2005. Optimization of initial centroids for k-means using simulated annealing. In Proc. Industrial Electronics Seminar (IES) 2005 (pp. 286-289).

[3] Ferguson, T.S., 1961. Rules for rejection of outliers. Revue de l'Institut International de Statistique, pp.29-43.

[4] Istas Pratomo, Eni Yusriani, Yoyon K. Suprapto. 2014. Klasifikasi Trafik Internet Menggunakan Metode Naive Bayes. Proceedings of the SISTI Seminar.

[5] Kumar, R., 2009, June. Mining web logs: applications and challenges. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 3-4). ACM.

[6] Rousseeuw, P.J. and Kaufman, L., 1990. Finding Groups in Data. Wiley Online Library.

[7] Lu, B. and Ju, F., 2012, August. An optimized genetic K-means clustering algorithm. In Computer Science and Information Processing (CSIP), 2012 International Conference on (pp. 1296-1299). IEEE.

[8] Prasetyo, E., 2014. Data mining mengolah data menjadi informasi menggunakan matlab. Yogyakarta: Andi Offset.

[9] Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of computational and applied mathematics, 20, pp.53-65.

[10] Santosa, B., 2007. Data mining terapan dengan matlab. Yogyakarta: Graha Ilmu.

[11] Suwirmayanti, P., Putra, I. and Kumara, I., 2014. OPTIMASI PUSAT CLUSTER K-PROTOTYPE DENGAN ALGORITMA GENETIKA. Majalah Ilmiah Teknologi Elektro, 13(2).

[12] Alfina, T., Santosa, B. and Barakbah, A.R., 2012. Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabungan Keduanya dalam Cluster Data (Studi Kasus: Problem Kerja Praktek Teknik Industri ITS). Jurnal Teknik ITS, 1(1), pp.A521-A525.

[13] Tan, P.N., Steinbach, M. and Kumar, V., 2013. Data mining cluster analysis: basic concepts and algorithms. Introduction to data mining.

[14] Yuhefizar, Budi S,I Ketut E, Yoyon K Suprapto, Two Level Clustering Approach for Data Quality Improvement in Web Usage Mining. Journal of Theoretical and Applied Information Technology. 2014:62(2):404-409.

[15] Zukhri, Z., 2014. Algoritma Genetika: Metode Komputasi Evolusioner untuk Menyelesaikan Masalah Optimasi.

BIODATA PENULIS

NUR ULFATUR ROIHA, lahir pada 26 Pebruari 1980 di Surabaya, Jawa Timur. Anak keempat dari tujuh bersaudara. Dibesarkan dalam keluarga sederhana di Surabaya Barat. Alhamdulillah, penulis dapat mengenyam pendidikan formal di SDN Tandes Kidul I, lulus tahun 1993. Selanjutnya meneruskan pendidikan di SMPN 2 Surabaya, lulus tahun 1996 dan melanjutkan ke SMUN 5 Surabaya, lulus tahun 1999.

Setelah mengikuti ujian UMPTN, penulis diterima di Universitas Airlangga Surabaya, namun hanya ditempuh selama satu tahun. Kecintaan terhadap dunia komputer membuat penulis mencoba mengikuti UMPTN lagi pada tahun 2000. Alhamdulillah diterima di S1 Jurusan Teknik Informatika, Institut Teknologi Sepuluh Nopember, lulus tahun 2005.

Selepas meraih sarjana komputer, penulis bekerja sebagai tenaga programer di software house dan pada tahun 2008 mengikuti ujian PNS, Alhamdulillah diterima dan ditempatkan di Dinas Komunikasi dan Informatika Kota Surabaya. Penulis berkarir disana hingga saat ini.

Alhamdulillah, penulis berkesempatan melanjutkan Studi Magister pada tahun 2014 di Jurusan Teknik Elektro, Institut Teknologi Sepuluh Nopember, yang ditempuh selama 5 semester dan lulus tahun 2017.

Alhamdulillah, penulis melepas masa lajang pada tanggal 14 September 2001 dan mendapatkan pria “Lelananging Jagad” yang rumahnya bertetangga dengan penulis. Sampai saat ini, Alhamdulillah, Allah berkenan menitipkan putri shalihah yang lahir pada tanggal 26 April 2006.

Penulis dapat dihubungi di nomor 087751652942, alamat e-mail: nurulfa998@gmail.com.